American Review of Public Administration
1–23

DOI: 10.1177/0275074014524299
arp.sagepub.com
SAGE

A Bayesian Approach to Measurement Bias in Networking Studies

Ling Zhu[1], Scott E. Robinson[2], and René Torenvlied[3]

Abstract

The study of managerial networking has been growing in the field of public administration; a field that analyzes how managers in open system organizations interact with different external actors and organizations. Coincident with this interest in managerial networking is the use of self-reported survey data to measure managerial behavior in building and maintaining networks. One predominant approach is to generate factor indices of networking activity from ordinal scales. However, when public managers answer survey questions with ordinal scales to describe their networking activities, the answers may be subject to various response biases. Consequently, the use of factor indices may lead to biased measurements that misrepresent managerial networking. As an alternative, we build on studies that apply the item response theory (IRT) as a measurement strategy and propose a Bayesian alternative. To tap managers' latent effort put in networking activity, the Bayesian Generalized Partial Credit Model allows us to select a one-dimensional networking scale from multiple ordinal survey items. Using 12 such items in a mail survey of nearly 1,000 American hospital managers, we demonstrate the advantage of using the Bayesian IRT model over factor-analytic models in a substantive test of how managerial networking affects organizational performance.

Keywords

managerial networking, measurement bias, Bayesian IRT

Introduction

The field of public management is flourishing, with theoretical-empirical research contributing to diverse areas of study, such as public sector motivation, red tape, collaborative management, performance management, or networking. Empirical evidence in the field is rapidly accumulating, relying on equally diverse methods, such as (field) experiments, content analysis, comparative case studies, or survey research. In surveys, we aim to measure traits of our respondents, tapping important information about the public management phenomena under study or about factors that partially explain these phenomena. Measures for public sector motivation, for

[1]University of Houston, Houston, TX, USA
[2]The University of Oklahoma, Norman, OK, USA
[3]Twente University, Enschede, The Netherlands

Corresponding Author:
Ling Zhu, Department of Political Science, University of Houston, 436 Philip G. Hoffman Hall, Houston, TX 77204-3011, USA.
Email: lzhu4@central.uh.edu

example, combine information from different items that are constructed as ordinal (Likert-type) scales. These scales measure the level of agreement of respondents with a number of statements about various aspects of their motivation, varying from 1 = *strongly disagree* to 5 = *strongly agree* (Kim, 2011; Perry, 1996). Another example, elaborated in the present article, is the measurement of the intensity of managerial networking. Managerial networking is the extent to which managers maintain relations with various types of external organizations and actors, for example, suppliers, stakeholders, clients, alliance partners, regulatory agencies, or political actors and institutions. Networking activity is conceptualized as the contact frequency of relations that (high ranking) managers maintain with external actors and organizations. Each item is a type of external actor or organization, while the scale taps the contact frequency with these organizations in ordinal categories (varying from 1 = *never* to 6 = *daily*).[1]

The dominant approach to measurement in public management research is to apply factor analysis techniques to self-reported ordinal survey data. Such an approach implicitly assumes that different survey items share the same underlying distribution. For studies of managerial networking, the theoretical literature does not offer much guidance on the dimensionality of networking activities, with Thomson, Perry, and Miller (2009) as a notable exception. Studies on the Texas school district data repeatedly report the existence of a single, continuous factor underlying these contact frequencies, which is positively associated with various measures of school district performance (Meier & O'Toole, 2001). Yet, although the results of empirical studies into managerial networking seem promising, some cautionary remarks must be made when it comes to measurement theory. Too hastily the field of public administration has rushed into performance appraisal and performance management without paying proper attention to measurement theory and methodology (Meier & O'Toole, 2013; Torenvlied & Akkerman, 2012). The quality of our measurement critically affects our ability to test causal relationships, make proper statistical inferences, and, more generally, our understanding of administrative behavior. Improving the measurement procedures that map administrative and managerial behaviors is critical to the building of knowledge in our field.

A common problem with the dominant approach to measurement is that the latent traits, derived from factor analysis, which scholars intend to measure, do not have well-established units of measurement. More specifically, the application of factor-analytic methods to ordinal data poses two major challenges to measurement validity. The first challenge is the *compatibility issue*, meaning that respondents may understand the same question in different ways (Brady, 1985; King, Murry, Salomon, & Tandon, 2004; King & Wand, 2007). The potential heterogeneity across respondents is, "a typical source of variation in response data that needs to be accounted for in a statistical response model" (Fox, 2010, p. 3). Survey items on managerial networking are extremely vulnerable to this validity threat in that both managers' personal traits and their organizational contexts might induce a specific networking activity and affect its frequency. A simple index without accounting for the cross respondent heterogeneity will be a biased measure, which is not compatible across all respondents.

The second challenge is that perceptual survey items may induce a *social desirability* bias, meaning that respondents have a tendency to answer questions in a socially desirable way, thus producing a "common source bias" (Donaldson & Grant-Vallone, 2002; Doty & Glick, 1998; Graham & Collins, 1991; Podsakoff & Organ, 1986; Schwarz, 1999). Both the factor-analytic technique and the non-parametric cumulative scaling technique assume that each behavior item is correctly measured. In other words, measurement scales are produced with the assumption that managers truthfully report their activities of managerial networking. Recent studies show, however, that self-reported survey items are vulnerable to various reporting bias. Meier and O'Toole (2013) demonstrate that managers might over- or underreport particular networking activities due to the social desirability of the survey item. Henry, Lubell, and McCoy (2012) find that self-reported survey items are likely to have recall bias, namely, respondents are unlikely to remember their entire list of network nodes and report accurately how much they networked with each

node. These most recent studies, nevertheless, are long on description and short on prescription of how to quantify managerial networking, accounting for measurement bias embedded in the survey data.

Two recent contributions in the subfield of managerial networking, indeed, show how sensitive our analyses can be to the measurement model applied. Robinson and Gaddis (2012) compare different measurements of collaboration. They analyze a survey of school district superintendents following Hurricane Katrina to assess postdisaster collaborative activity. They compare various approaches to measuring collaborative activity and report that varying questions unsurprisingly led to varying distributions of collaborative activity, even though these varying distributions were captured by a single underlying factor. In a recent study, Torenvlied, Akkerman, Meier, and O'Toole (2013) introduce "item response theory" (IRT) as an alternative to factor analysis to arrive at networking activity scales, mapping different ordinal contact frequency items to cumulative scales. The IRT approach reveals that the contact frequencies of Texas school district superintendents cannot be mapped on a single networking activity dimension, but on several dimensions, each related to a specific source of support from the environment of their school districts.

The present article offers two main contributions to the existing literature on measurement in public management, more in particular on managerial networking. The first main contribution is to introduce an approach which addresses the important, yet overlooked, problem of measurement bias in the study of networking activities. Drawing from the IRT literature (Embretson & Reise, 2000; van der Linden & Hambelton, 1996), including non-parametric Mokken Scale analysis (Mokken, 1971; van Schuur, 2003, 2011), we propose a Bayesian variant of the IRT approach to infer the latent dimension(s) of managerial networking. Bayesian IRT models have been introduced in other fields such as education (e.g., Lord, 1986; Patz & Junker, 1999) and psychology (Fox & Glas, 2003). The Bayesian approach to measurement conceptualizes the effort level of managerial networking as a latent variable, which is not directly observable. Theoretically, managerial networking activities are, in the first place, driven by time-budget constraints which translate into the difficulty managers have to concentrate their time-allocation into one networking node. Managerial networking activities are, in the second place, driven by time-allocation decisions based on returns to investment in different networking dimensions; that is, the discrimination of one networking node from all the other nodes (Akkerman & Torenvlied, 2011; Torenvlied et al., 2013).

A Bayesian model, furthermore, deals with uncertainty in parameters caused by response bias. We argue that the Bayesian approach of measurement is more attractive than the factor-analytic methods because it relaxes the assumption that responses to each networking item are treated with equal weights. The Bayesian approach also accounts for differential response difficulties for different respondents by explicitly modeling differential item functioning (DIF; Embretson & Reise, 2000). Finally, the Bayesian IRT approach extends the non-Bayesian IRT models in that it is exible to handle cross-item heterogeneity and uncertainty in parameters. Thus, the approach in the present article improves measurement reliability (Fox, 2005, 2010; Jackman, 2009).

The second main contribution of the present article is that we apply our analyses in the field of professional health care organizations (American hospitals), using 12 items on networking activities from a large survey. We illustrate the statistical procedure of computing the Bayesian IRT measure. We demonstrate that the Bayesian approach to measurement helps to reduce measurement bias associated with the classic factor-analytic methods and is particularly useful when the number of observations is relatively small.

Networking as a Latent Variable

We propose a Bayesian Item Response approach to model ordinal survey items of managerial networking. Our approach builds on the IRT and the cumulative scaling technique, but explicitly

accounts for uncertainty in parameters. Bayesian inference is then used to improve measurement reliability.

Ordinal Response Data and the IRT Approach

The IRT was initially developed in psychological measurement and educational testing (Lord & Novick, 1968). Despite their different mathematical forms, IRT models conceptualizes that the probability of a particular response to a survey item is a function of the respondents' and the survey items' characteristics (Hemker, Sijtsma, & Molenaar, 1995). The IRT scaling approach differs from the classic reliability (factor) analysis. Factor (reliability) analysis assumes that all items have the same underlying frequency distributions, whereas IRT models explicitly account for heterogeneous item distributions (van Schuur, 2003).

In the context of analyzing dichotomous item responses,

> the answer a person gives to a question is interpreted as a dominance relationship between the person and the question (the item). The person dominates the item if she gives the positive response, and the item dominates the person if she gives the negative response. (van Schuur, 2011, p. 16)

With a large number of survey items and respondents, one can analyze four types of dyadic dominance relationships: item-to-subject, subject-to-item, item-to-item, and subject-to-subject. Without measurement errors, this scaling approach can produce a transitive rank-order among all items based on the "difficulty" parameter and a transitive rank-order among all items based on the "discrimination" parameter. The latent dimension of ability (or other latent traits, such as effort, ideology, attitude, etc.), therefore, is inferred and scaled based on the cumulative probability of giving a positive response identified by the difficulty and discrimination parameter.

Both Rasch (1960) models and the Mokken (1971) scale, for example, define the probability of observing a positive response to a survey (or test) item based on the respondents' latent traits and one or more item parameters. Rasch models, known as one-parameter logistic IRT models, only include item difficulty as a determinant of observed responses. Mokken scaling adds a specific discrimination parameter for each item and make IRT models more flexible for fitting different empirical datasets. Hence, Rasch models can be viewed as a unique case of the Mokken scale, where the item discrimination parameter is restricted to be a constant.

When applying the IRT (Mokken) approach to ordinal responses, the model essentially becomes a cumulative rating scale model or a partial credit model (Anderson, 1997; Masters, 1988; Samijima, 1969). Ordinal response items in networking studies often generate a scale based on ordered categories. The survey question may ask a respondent using letter grades, A, B, C, and D (ordered from low to high), to indicate the effort of his networking activities. A survey item may also ask a respondent the frequency of her networking activities, rendering a scale, ordered as *never, yearly, monthly, weekly, daily*, and *more than once per day*. Attitude items, in addition, normally produce an ordered scale ranked by *strongly disagree, disagree, agree*, and *strongly agree*. Compared with the dichotomous survey items, ordinal items, or more generally, polytomous items increase the statistical information (Ostini & Nering, 2006) that researchers can use to infer managerial ability or effort of networking.

Let the latent dimension of networking effort for i managers be denoted by θ_i, one can use j items to denote different networking activities. Each survey item j then will produce a stack of item-subject matrixes, and have a density distribution based on all the reported choices for that particular item. In turn, one can then use the stack of survey items to estimate the latent effort of managerial networking. Specifically, for a survey item j (networking node), the probability of a manager i to report a specific response k in the ordinal k-choice category is defined by the difference between the probability of responding in (or above) the previous choice $k-1$ and the probability of responding in (or above) the choice, k (Fox, 2010).

$$P\left(Y_{i,j}=k\mid\theta_i,\alpha_j,\beta_j\right)=P\left(Y_{i,j}\geq(k-1)\mid\theta_i,\alpha_j,\beta_j\right)-P\left(Y_{i,j}\geq k\mid\theta_i,\alpha_j,\beta_j\right). \quad (1)$$

The cumulative probability of choosing the lowest category and above is assumed to be 1 and the probability of choosing above the highest category is 0. As for a k-choice ordered scale, there are $(k-1)$ cut points to define the ordinal scale. Hence, Equation 1 specifies the two probability values with respect to cut point $k-1$, and k. In Equation 1, α_j and β_j are the discrimination and difficulty parameter, respectively. θ_i is the latent dimension at interest. The two probabilities in Equation 1, further, can mathematically be represented by the logistic cumulative distribution function (Bradlow & Zaslavsky, 1999; Fox, 2010; Fumiko, 2008).

$$\begin{aligned}P\left(Y_{i,j}=k\mid\theta_i,\alpha_j,\beta_j\right)&=P\left(Y_{i,j}\geq(k-1)\mid\theta_i,\alpha_j,\beta_j\right)-P\left(Y_{i,j}\geq k\mid\theta_i,\alpha_j,\beta_j\right)\\&=\frac{exp\left(\alpha_j\left(\theta_i-\beta_{j,k-1}\right)\right)}{1+exp\left(\alpha_j\left(\theta_i-\beta_{j,k-1}\right)\right)}-\frac{exp\left(\alpha_j\left(\theta_i-\beta_{j,k}\right)\right)}{1+exp\left(\alpha_j\left(\theta_i-\beta_{j,k}\right)\right)}\quad. \quad (2)\end{aligned}$$

Considering Uncertainty in Parameters: A Bayesian Approach

The Bayesian IRT approach we propose shares the same theoretical foundation with non-Bayesian IRT models, but explicitly acknowledge uncertainty in parameters. Reporting bias and the issue of cross-item compatibility can both change the perfect rank-order estimated based on the difficulty and discrimination parameter. The IRT conceptualizes measurement bias as the violation of a transitive relationship between items and subjects. Bias, counted as "reporting errors" (or "wrong answers"), thus, can be estimated through evaluating the homogeneity of individual items based on pairwise information extracted from each item pairs (van Schuur, 2011). Figure 1 illustrates how the pairwise association may look like between two ordinal survey items. In Figure 1, both items are coded based on a 1-to-6 ordinal scale, and each circle indicates the frequency of all possible dyadic values, whereby a large circle refers to high frequency and small circle refers to low frequency. Ideally, if Items 1 and 2 are created to measure the same latent construct, the association between the two items should only be caused by their common associations with the latent factor. Local independence is a key underlying assumption of latent variable models, which requires that the observed items are independent from each other given an individuals' score on the latent variable. When this assumption holds, the tabulation of the two items would be a more cumulative pattern, such as a lower or upper triangle (Torenvlied et al., 2013). The pattern in Figure 1 demonstrates the violation of this local independence assumption. From the tabulation of all 36 possible pairwise values between the two items, three pairs (25, 26, and 62) are not observed. Most observed information concentrates on the dyadic combinations, when both Items 1 and 2 take values greater than 2. When the local independence assumption is violated, the estimated latent variable does not fully account for the associations between the observed items. Moreover, Figure 1 shows that individual respondents are more likely to check high values for Item 1 than for Item 2, indicating different item difficulty levels of the two items. Classic factor-analytic models do not fully account for heterogeneity in item difficulty.

The key challenge to measure a latent trait based on observed survey items is that we do not know the true measurement score of the latent construct. We empirically estimate the latent traits and measurement errors simultaneously. In standard non-parametric IRT, this procedure results in point-estimates for the discrimination and difficulty parameter Thus, when observing data as shown in Figure 1, it is difficult to model the uncertainty in parameters. The Bayesian IRT approach becomes particularly appealing to handle such measurement issue, because neither the latent traits, nor the item parameters need to be treated as deterministic.

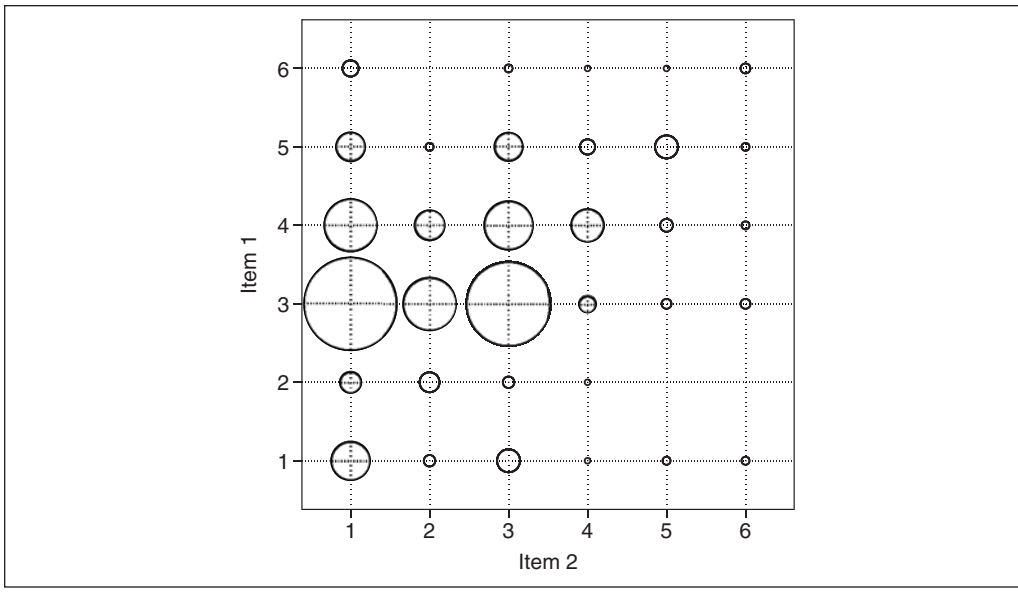


Figure 1. An illustration of the association between an item pair.

Re-arranging the likelihood function expressed by Equation 2, we define the probability of manager i reporting a particular choice k_j for survey item j can be expressed as a cumulative logistic density below the latent cut point $\tau_{k,j}$ associated with choice k_j.

$$P\left[Y_{ij} = k_j\right] = P\left[\alpha_j\left(\theta_i - \beta_j\right) \le k_j\right] = 1 - Z\left(\tau_{k,j} - F\left(\theta_i, \alpha_j, \beta_j\right)\right). \tag{3}$$

To simplify the mathematic notation, in Equation 3, we use Z(.) to denote the cumulative density function, and $F(\theta_i, \alpha_j, \beta_j)$ to denote the linear link function for $Y_{i,j}$ defined by latent trait θ_i, the discrimination parameter α_j, the difficulty parameter β_j. $\tau_{k,j}$ is the underlying cut points specific to the categorical values k associated with j survey items.

In the context of estimating a full Bayesian model, all the parameters in $F(\cdot)$ are estimated quantities, thus, carry measurement uncertainty in parameters. Specifically, the density function of the latent managerial networking dimension and the measurement bias in survey items are all unknown. We are interested in estimating their joint densities based on the assumed likelihood function, the prior information on their distribution, and the observed survey response data. The term $\pi(\theta, \alpha, \beta)$ denotes the joint posterior density that we want to infer, whereby θ refers to the latent trait of managerial networking, α refers to the vector of discrimination parameters, and β refers to the vector of difficulty parameters. The marginal posterior density is learned based on using observed data to update prior assumptions about how these unknown parameters are distributed. Formally, the posterior density is given as

$$\pi\left(\theta, \alpha, \beta \mid Y\right) \propto L\left(Y \mid \theta, \alpha, \beta\right) p\left(\theta\right) p\left(\alpha\right) p\left(\beta\right). \tag{4}$$

Using the Bayesian approach, one not only can obtain the point estimation (e.g., posterior means) of managerial networking (θ_i) and the difficulty/discrimination parameter, but also can

Table 1. Comparison of the Assumptions of Three Implicit Measurement Strategies.

Assumption	Factor analysis	IRT	Bayesian IRT
Local independence	Yes	Yes	No
Homogeneous item difficulty	Yes	No	No

Note. IRT = item response theory.

estimate the Bayesian credible intervals for each estimated parameter (e.g., see Treier & Jackman, 2008). In the next section, we apply the proposed Bayesian IRT model using items from a survey of professional health care organizations. We compare measurement reliability of networking scales across three different measurement approaches of managerial networking: (a) a principal factor index, (b) a non-Bayesian IRT index, and (c) a Bayesian IRT index. The three networking indices are, furthermore, compared with respect to their predictive validity in accounting for organizational performance.[2]

An Empirical Application: Comparing Three Measures of Managerial Networking in American Hospitals

Data

The empirical data, which we use to illustrate the Bayesian IRT approach to measurement, are drawn from a mail survey of 6,000+ professional health care organizations, including general hospitals, specialized hospitals, mental health clinics, children's hospitals, university-owned medical centers, rehabilitation centers, and acute long-term care organizations (Johansen & Zhu, 2014). The survey was administrated from December 2010 to March 2011 and generated 1,004 responses, rendering an overall response rate of 15.85%. Although the overall response rate is not high, it is quite comparable with other large-scale surveys on American hospitals.[3] We perform a logistic regression analysis to check if response rates vary substantially by organizational type, region, state, service type, and organizational size. Our sample does not vary substantially along all these indicators excepting organizational type.

The survey uses a name-roster method, providing top-level hospital administrators an elaborate list of networking nodes. To measure managerial networking, we use 12 nodes from the list that capture different types of external partners (see Figure 2).[4] Each survey item is measured by a 1-to-6 ordered scale ranking the frequency of interactions between a manager and her networking partners. The order scale is defined as 1 = *never*, 2 = *yearly*, 3 = *monthly*, 4 = *weekly*, 5 = *more than once a week*, and 6 = *daily*. This approach is analogous to that used by Agranoff and McGuire (1999), Meier and O'Toole (2001), and others in defining managerial networking as an activity that crosses organizational boundaries. In other words, this survey design focuses on measuring the behavioral traits of managerial networking, not the structural characteristics of networks.

In our sample, there are 338 government-owned (primarily non-federal hospitals), 480 non-profits, and 195 for-profits hospitals. Given that survey response rates vary across organizational ownership (see Table 2), we estimate a logistic regression model predicting survey response rate. We include hospital service type, geographic location (state and American Hospital Association [AHA] service area), ownership, and hospital size as predictors in the logistic regression model. As such, we examine whether these factors affect survey response rate, if so, whether the size of non-response bias is troublesome. Table 3 reports detailed information about response rates by hospital ownership and the logistic regression model. Although survey response rates vary across hospital ownership types, the non-response bias introduced by ownership is not very large, with a negative coefficient of −0.038. Geographic location and hospital size both significantly predict

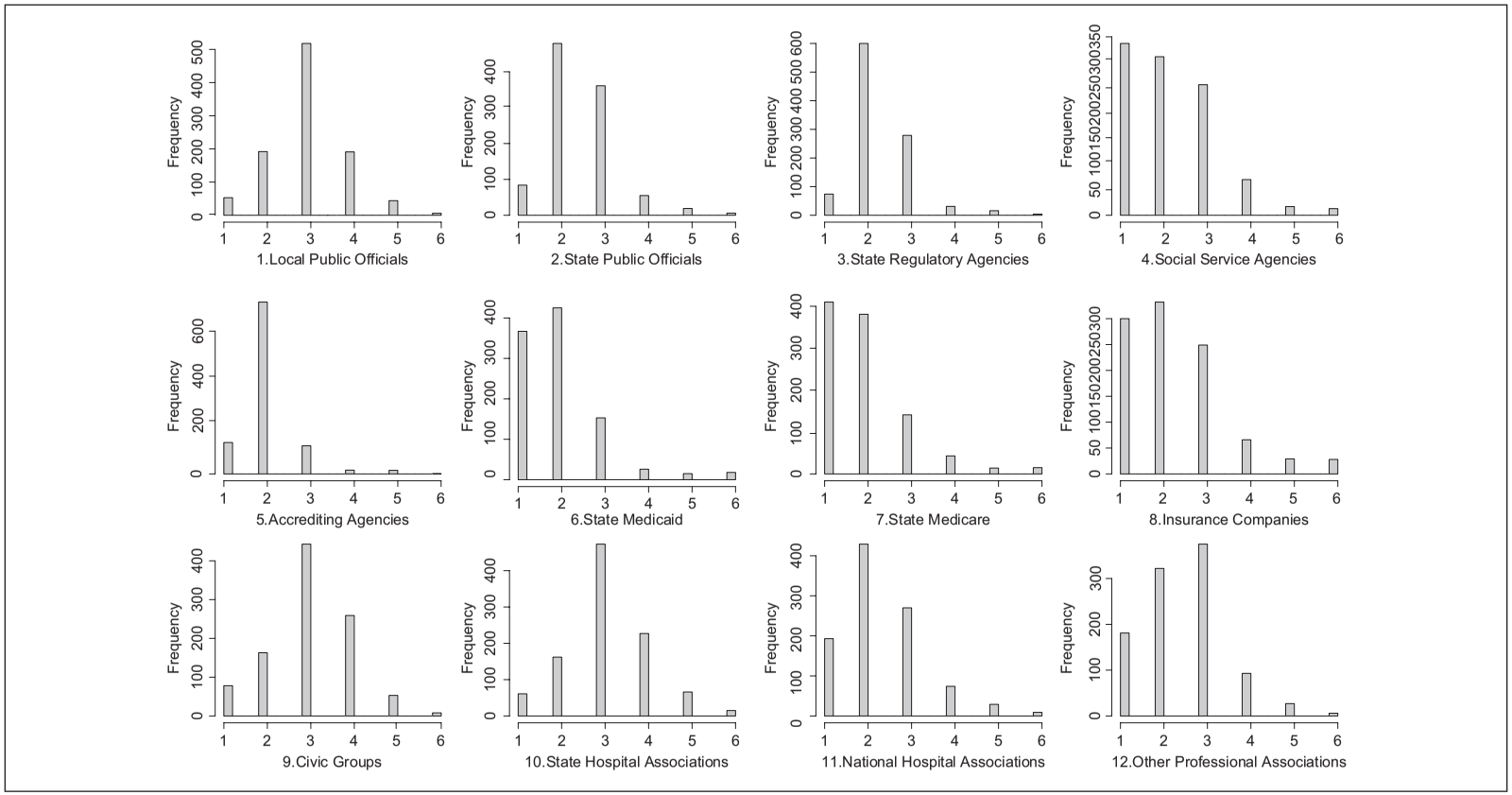


Figure 2. Frequencies of managerial networking with 20 external partners (1 = *never* and 6 = *daily*).

Table 2. Survey Response Rates by Organizational Type.

Owned by	AHA code	No response	Response	Total	Rate (%)
State	12	266	57	323	17.65
County	13	281	90	371	24.26
City	14	72	30	102	29.41
City-county	15	25	4	29	13.79
Hospital district	16	416	135	551	24.50
State/local govt. overall		1,060	316	1,376	22.97
Air force	41	9	1	10	10.00
Army	42	20	2	22	9.09
Navy	43	11	2	13	15.38
PHS	44	9	1	10	10.00
Veteran service	45	120	15	135	11.11
PHS Indian service	47	23	1	24	4.17
DOJ	48	2	0	2	0.00
Federal overall		194	22	216	10.19
Church	21	472	74	546	13.55
Other	23	2,189	406	2,595	15.65
Non-profit overall		2,661	480	3,141	15.38
Individual	31	25	5	30	16.67
Partnership	32	235	50	285	17.54
Corporation	33	1,155	131	1,286	10.19
Private overall		1,415	186	1,601	11.62
Total		5,330	1,004	6,334	15.85

Note. AHA = American Hospitals Association; PHS = public health service; DOJ=Department of Justice.

Table 3. Logistic Regression Predicting Survey Responses (1 = *Response*, 0 = *Non-response*).

Variable	Coefficient	*SE*
Service type	−0.002	0.002
State code	0.006**	0.001
AHA area	0.00002	0.0001
Ownership	−0.038**	0.005
Hospital size	−0.0003**	0.00005
Intercept	−0.883**	0.176
N	6,334	

Note. AHA = American Hospitals Association.
Significance levels: †10%. *5%. **1%.

survey response rates, but their substantive impacts are also minimal. In this section, in which we analyze the relationship between managerial networking and hospital performance, we control for different organizational types.

Three Measures of Managerial Networking

We use three different measurement approaches to estimate the latent dimension of managerial networking: (a) factor analysis, (b) a Generalized Partial Credit Model (GPCM) in an empirical Bayes context, which essentially applies the Mokken scale analysis to polytomous survey

responses, and (c) a full Bayesian GPCM. Using the *stats* package in *R*, we estimate a factor index using the principal factor analysis and recover managerial networking by predicting the first factor score. The empirical Bayes IRT scale is estimated by fitting a GPCM using the *ltm* package in *R* in a non-Bayesian setup. The latent scale of managerial networking is predicted using the *empirical Bayes* method after fitting the IRT model using the gpcm( ) function (Rizopoulos, 2006).[5] As for estimating the full Bayesian IRT scale, we use *R* and *JAGS* to implement the Bayesian GPCM. The Bayesian GPCM returns a full set of discrimination parameters (α_j), difficulty parameters ($\beta_{j,k}$),[6] and recovers the marginal posterior distributions of the latent dimension of managerial networking for each hospital. Using the full set of posterior distributions, θ_i, we infer the Bayesian managerial networking index based on posterior means draw from θ_i.

Because prior specification can affect the results of Bayesian inferences (Gelman, Carlin, Stern, & Rubin, 2003; Gill, 2008), we follow Gelman (2006), Li and Baser (2012), and Treier and Jackman (2008) and specify diffuse priors with zero-mean and large-size variance to all parameters.

$$\begin{aligned} \alpha &\sim N(0, 0.001) I(0,) \\ \beta &\sim N(0, 0.001) I(-3.5, 3.5). \\ \theta &\sim N(\mu, \sigma) \end{aligned} \tag{5}$$

We specify α as a normal distribution with mean 0 and precision 0.001. We also set α to take positive values. The truncation is chosen because α is often positive and near 1 in IRT applications. The noninformative prior specifies a large-size variance and a zero-mean, as such, it does not inform the model about the specific value of α. Similarly, we specify a noninformative prior for β, with mean 0 and precision 0.001. Based on the results of the empirical Bayes GPCM, we place most parameter values in the range (−3.5, 3.5). We did not set θ to be a standard normal distribution. Instead, we adopt the strategy of prior specification in Li and Baser (2012), and set the hyper-prior, μ (underlying mean for the latent networking dimension, θ) to be a normal distribution with mean 0 and precision 0.001. The hyper-prior, σ is defined as a uniform distribution.[7] Similar to the identification strategy used in Treier and Jackman (2008), we center each estimated θ_i to have zero mean and scale them based on the estimated σ. In other words, we impose the rescaling to the Markov Chain Monte Carlo (MCMC) outputs by each iteration. Accordingly, we rescale α_j and $\beta_{j,k}$ using the same strategy.[8]

After estimating the three managerial networking indices, we compare them by evaluating measurement reliability and predictive validity. We then use Monte Carlo experiments to assess bias-reduction moving from the factor index to the full Bayesian IRT index.

Measurement reliability. Table 4 reports results from the principal factor analysis, which shows that the 12 managerial networking items do not load on one single factor but on three factors. The first factor is driven by contact with financial stakeholders, such as state Medicaid, Medicare, and insurance complains. The second factor is driven by contact with political officials, and the third factor reflects contact with peer organizations. The factor index of managerial networking is predicted based on the first dimension. As a result, the index performs poorly in recovering data information efficiently. Figure 3 plots factor loadings along the first two dimensions. It shows that only about 34% information is retained in the networking index based on the first dimension.

Ignoring cross-item and cross-choice heterogeneity, the networking measure produced by factor analysis is influenced by survey items that have distributions strongly skewed toward the

Table 4. Managerial Networking: Factor Loadings Based on Principal Component Analysis.

Networking node	Factor 1	Factor 2	Factor 3
Local public officials	0.081	**0.543**	0.240
State public officials	0.163	**0.773**	0.162
State regulatory agencies	0.418	**0.592**	0.156
State Medicaid	**0.722**	0.270	0.139
State Medicare	**0.830**	0.049	0.135
Accrediting agencies	**0.357**	0.290	0.182
Social service agencies	**0.381**	0.260	0.068
Civic groups	0.016	**0.337**	0.318
Insurance companies	**0.625**	0.015	0.105
State hospital associations	0.086	0.130	**0.621**
National hospital associations	0.209	0.152	**0.781**
Other professional associations	0.167	0.248	**0.492**
Sum of squared loadings	2.160	1.685	1.532
Proportion variance	0.18	0.140	0.128
Cumulative variance	0.18	0.320	0.448

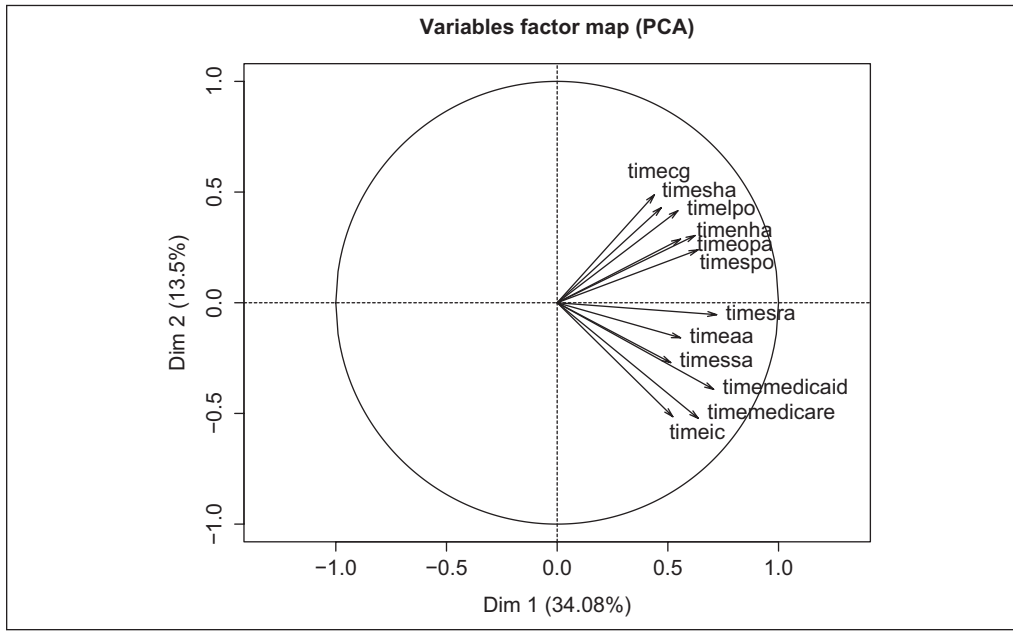


Figure 3. Plot factor loadings along the first two dimensions (based on principal component analysis).

low-frequency end. The non-Bayesian IRT model improves the factor index by making explicit assumptions about how each networking item and networking choices within each networking item contributes to the latent networking index. The issue of intra-item and intra-choice heterogeneity, however, still leads to some issues in parameter estimation when fitting a non-Bayesian GPCM. Figures 4 and 5 illustrate item information curve (IIC) and item characteristic curve (ICC) by each choice category. IIC indicates how an item is located on the latent networking scale, θ_i, where it is the best at differentiating managers' responses. ICC captures the relationship between a manager's response to a survey item and her score on the latent networking scale,

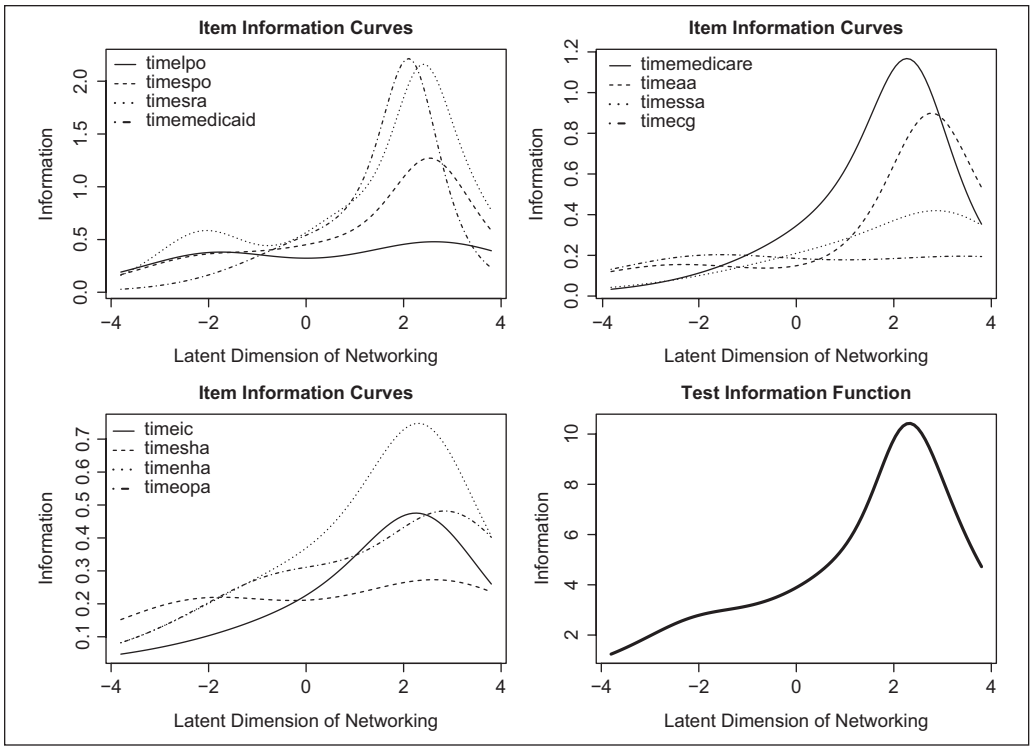


Figure 4. Item information characteristic curves based on the non-Bayesian Generalized Partial Credit Model.

θ_i. Overall, the non-Bayesian IRT model recovers 83.75% information in the estimated latent dimension of networking. The estimated scale, however, is skewed toward the high-end (see the *Test Information Curve* in Figure 4). The estimated latent scale recovers 61.03% information for values between 0 and 4 (high level of networking), but only 22.72% information is recovered for values between −4 and 0 (low-level of networking).

It is evident that the issues of inter-item and inter-choice heterogeneity still lead to problems in parameter estimation when fitting a non-Bayesian IRT GPCM. Comparing the by-item and overall IIC with the posterior mean distribution inferred from the Bayesian IRT model, the measurement scale returned by the Bayesian analysis converges to a normal distribution, whereby comparable information is measured for values below and above. Overall, the two IRT indices (empirical Bayes GPCM and full Bayesian GPCM) produce comparable estimation of item discrimination parameters and item difficulty parameters (see Statistical Appendix, Tables 1 and 3).

Figure 6 further compares the two IRT indices with the factor index. Figure 6 shows that the overall measurement agreement among the three networking indices is high, with pairwise correlations all being greater than 0.9. The two IRT indices, however, perform better than the factor index to score extremely low and high networking activities. Specifically, the factor index overestimates the networking scores in both the low- and high-end. In other words, compared with the two IRT indices, the factor index is more likely to assign relatively high networking scores to respondents (i.e., managers). The two IRT indices, in addition, produce comparable results (near 1 Pearson correlation), but the Bayesian IRT approach is superior with respect to estimating the networking scores for cases with extreme values. It is important to recall that many of the

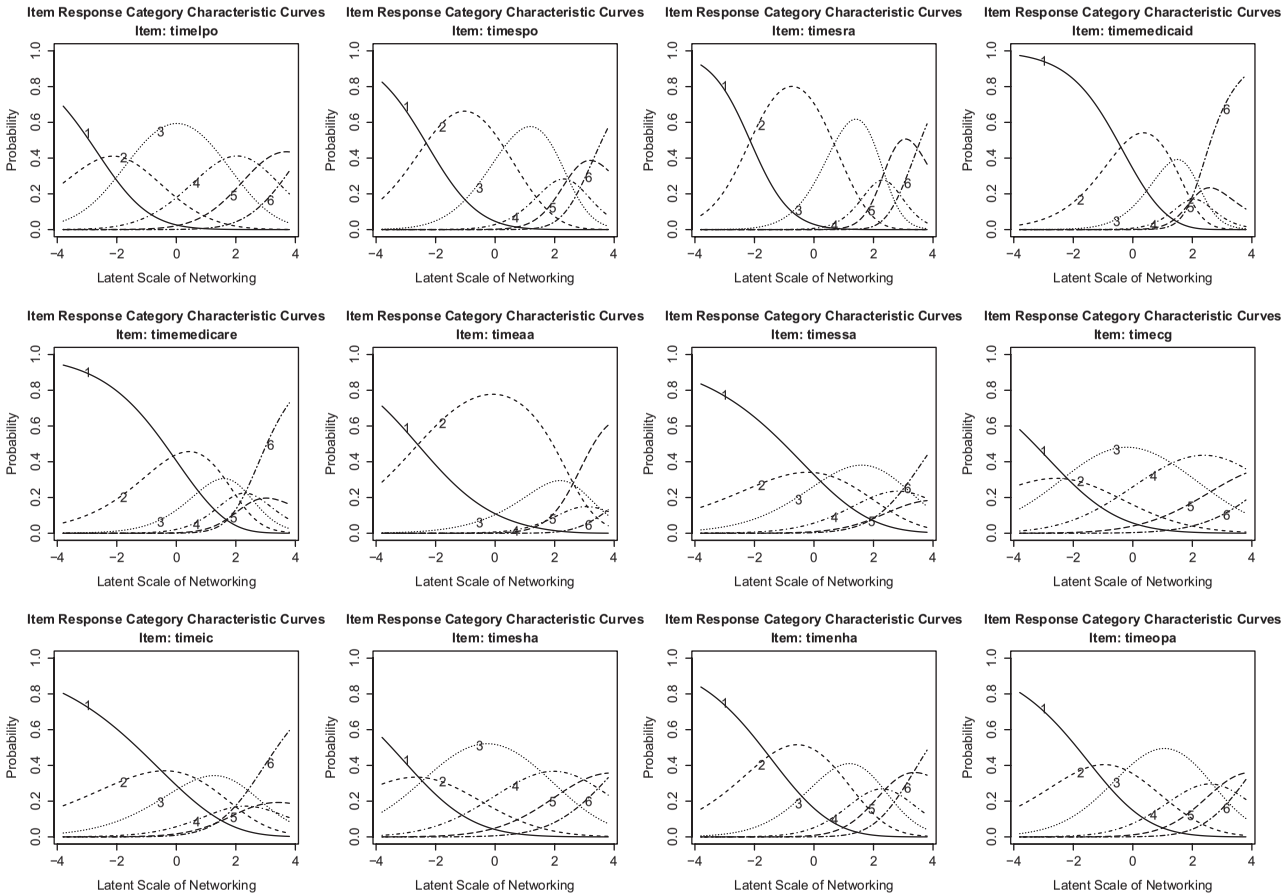


Figure 5. Item category characteristic curves based on the non-Bayesian Generalized Partial Credit Model (Category 1 = *never*, Category 2 = *yearly*, Category 3 = *monthly*, Category 4 = *weekly*, Category 5 = *more than weekly*, Category 6 = *daily*).

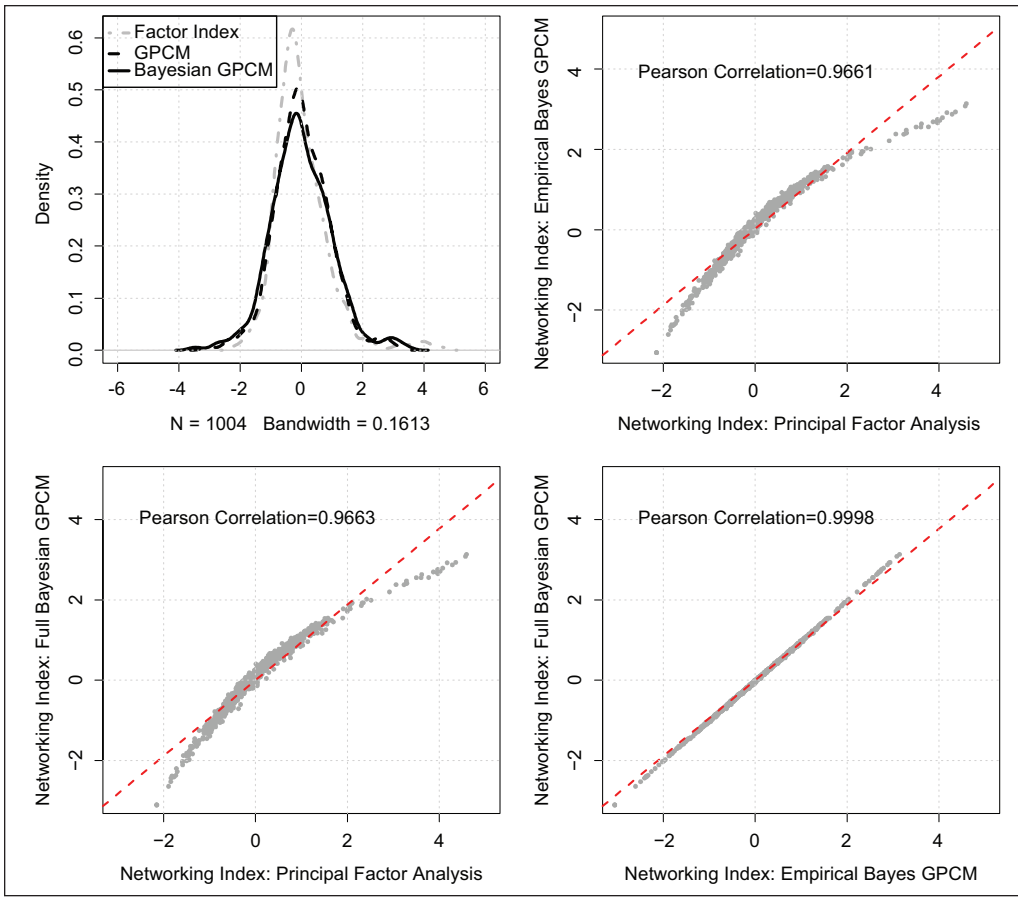


Figure 6. Measurement agreement among the three measures of managerial networking: Factor index, empirical Bayes GPCM index, and full Bayesian GPCM index.
Note. GPCM = Generalized Partial Credit Model.

individual node frequencies were strongly skewed, making sensitivity to extreme cases an important characteristic of any relevant technique.

Comparing predictive validity. To evaluate the predictive validity of the three networking measures, we estimate linear models for two hospital performance measures: total emergency room (ER) visits and total outpatient visits.[9] Following O'Toole and Meier (1999), we specify hospital performance O_i (measured by outputs) as a linear function of past performance in the previous year, managerial networking, and a vector of control variables reflecting managers' personal traits and available organizational resources.

The two performance measures (ER and outpatient visits) are drawn from the AHA FY 2010 Annual Hospital Survey Database. They are both scaled based on a log-transformation of the total counts, with greater measurement values representing more organizational outputs. Data on past performance are drawn from the AHA FY 2009 survey database. Also drawn from the AHA 2010 database are a battery of organizational resource variables, which are critical to service production. *Total Employee* is a 1-to-20 ordinal scale that measures the level of total full-time and part-time hospital employees. *Total Facility Expenditure* measures the level hospital facility expenses (including bad debt) for service production. It is also measured based on a 1-to-20

Table 5. The Association Between Hospital Performance (Emergency Room Visits) and the Three Managerial Networking Measures.

	Model 1		Model 2		Model 3	
Variable	Coefficient	*t* score	Coefficient	*t* score	Coefficient	*t* score
Managerial networking (factor index)	0.054*	2.15	—	—	—	—
Managerial networking (IRT)	—	—	0.077*	2.51	—	—
Managerial networking (Bayesian IRT)	—	—	—	—	0.079*	2.56
Total personnel	0.035	1.59	0.034	1.56	0.034	1.56
Total facility expenditure	−0.046	−0.29	−0.046	−0.29	−0.047	−0.29
Hospital size	−0.003	−0.05	−0.002	−0.03	−0.002	−0.03
Managerial tenure	0.003†	1.77	0.003†	1.75	0.003†	1.75
Public hospital	0.282**	3.25	0.284**	3.24	0.284**	3.24
Non-profit hospital	0.288**	3.80	0.294**	3.88	0.295**	3.88
In HMO network	−0.097	−1.56	−0.094	−1.52	−0.093	−1.52
In PPO network	0.137**	3.41	0.136**	3.46	0.136**	3.46
Total ER visits ($t - 1$)	0.957**	57.39	0.956**	56.82	0.956**	56.74
N	846		846		846	
R^2	.943		.943		.943	
RMSE	0.9307		0.9294		0.9293	

Note. IRT = item response theory; HMO = Health Maintenance Organization; PPO = Preferred Provider Organization; ER = emergency room; RMSE = root mean square error.
Significance levels: †10%. *5%. **1%.

ordinal scale, with greater measurement scores reflecting more expenses. *Hospital Size* measures the facility capacity of hospitals, which is a 1-to-8 ordinal scale, whereby "1" means 6 to 24 beds, and "8" means 500 or more beds in a hospital. *Managerial Tenure* is drawn from our hospital management survey and measures the total number of years a respondent has been a hospital manager. Because managers in different sectors place different priorities on the efficiency of service production, we also include two dummy variables for public and non-profit hospitals (Johansen & Zhu, 2014). Private hospitals are omitted as the baseline category. Because American hospitals are mostly connected within local providers' networks and rely on these formal networks to reach out to clients, we include two additional dummy variables to measure if a hospital is in a Health Maintenance Organization (HMO) network or in a Preferred Provider Organization (PPO) network.

For each performance measure, we estimate three models, using the same model specification, but alternating managerial networking variables. As such, the only difference across the three models is the managerial networking index, and we can compare coefficients across models to evaluate their predictive validity. We estimate all performance models using the ordinary least square (OLS) estimator with clustered standard errors by service types.[10] This is to control for heterogeneity across hospitals due to different service specialization. In addition, we include a full set of state-fixed effects in all models to control for unobserved heterogeneity.

Table 5 reports three models taking total ER visits as the performance measure. Model 1 uses the factor index of managerial networking, Model 2 includes the empirical Bayes IRT networking index, and Model 3 includes the full Bayesian IRT networking index. Models 1 to 3 all report positive associations between managerial networking and hospital performance measured by total ER visits. The estimated coefficient size, however, changes across the three models. Overall, models using the two IRT indices yield greater coefficients than the one reported in Model 1.

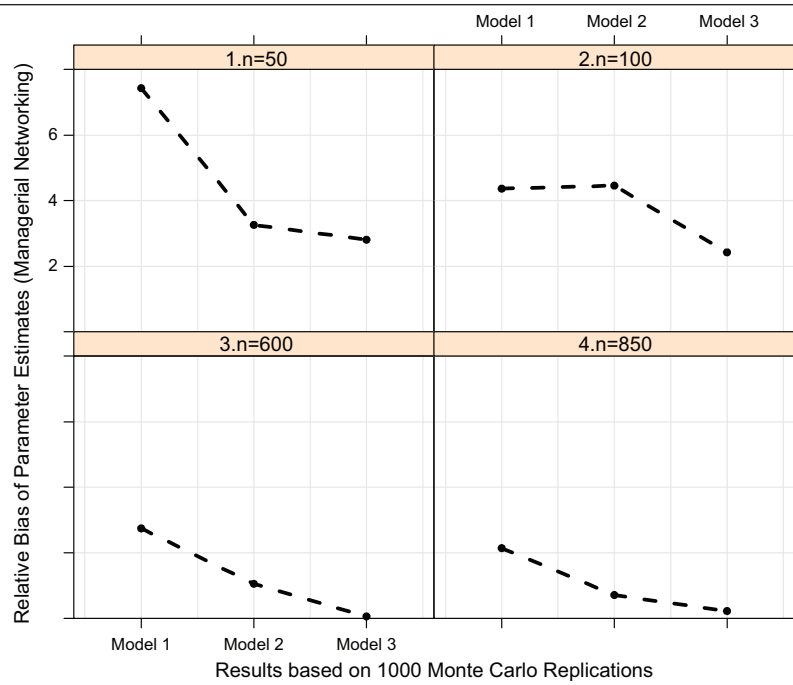
Table 6. The Association Between Hospital Performance (Outpatient Visits) and the Three Managerial Networking Measures.

	Model 4		Model 5		Model 6	
Variable	Coefficient	*t* score	Coefficient	*t* score	Coefficient	*t* score
Managerial networking (factor index)	0.095	1.41	—	—	—	—
Managerial networking (IRT)	—	—	0.120†	1.79	—	—
Managerial networking (Bayesian IRT)	—	—	—	—	0.124†	1.79
Total personnel	0.037	1.55	0.036	1.56	0.036	1.56
Total facility expenditure	0.022	0.16	0.022	0.17	0.022	0.16
Hospital size	−0.032	−0.48	−0.031	−0.46	−0.031	−0.46
Managerial tenure	0.005*	2.78	0.005*	2.78	0.005*	2.78
Public hospital	0.002	0.01	0.004	0.03	0.004	0.03
Non-profit hospital	−0.056	−0.02	0.003	0.02	0.003	0.02
In HMO network	0.158**	−1.01	−0.053	−0.97	−0.052	−0.96
In PPO network	0.158**	4.17	0.158**	4.14	0.158**	4.41
Total outpatient visits ($t - 1$)	0.963**	55.03	0.962**	54.41	0.962**	54.41
N	846		846		846	
R^2	.901		.902		.902	
RMSE	1.077		1.075		1.075	

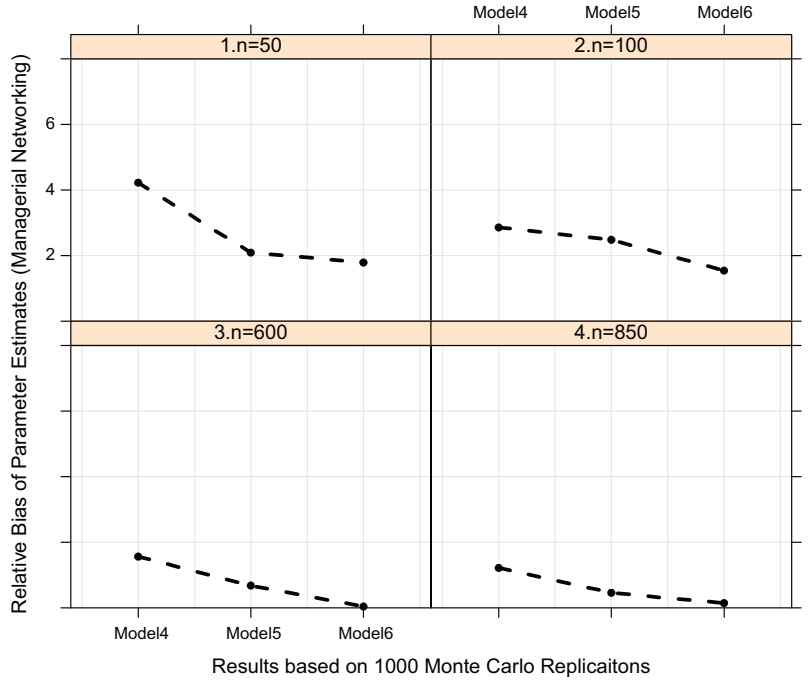
Note. IRT = item response theory; HMO = Health Maintenance Organization; PPO = Preferred Provider Organization; RMSE = root mean square error.
Significance levels: †10%. *5%. **1%.

Models 4 to 6 in Table 6 reveal a similar pattern. All three models report positive associations between managerial networking and hospital performance measured by total outpatient visits. Using the factor index, the estimated slope of networking management is 0.095 and statistically insignificant. Models 5 and 6 yield comparable slope coefficients, which are 0.120 and 0.124, respectively. Both coefficients, moreover, are statistically significant. Based on the point estimation of the slope coefficient, Model 4 seems to underestimate the impact of managerial networking on hospital performance. The comparison also suggests that the factor index may contain greater measurement error than the other two networking measures, because measurement error attenuates coefficient estimates. Last, comparing the three models in Tables 5 and 6, the two models using the full Bayesian IRT index produces the smallest root mean square error (RMSE) statistics (i.e., the best model fit by comparison).[11]

Reduced influence of measurement bias: Monte Carlo experiments. Tables 5 and 6 provide a straightforward comparison regarding how the point estimation of the slope coefficients would change by using the three different networking measures. All these parameter estimates, however, are computed based on one observed sample of managerial networking. Because we do not have repeated samples of Hospital managers, we use Monte Carlo simulation exercise to shed light on relative bias of parameter estimates for these competing models (Robert & Casella, 2010). For each set of three models, we conduct four Monte Carlo experiments with sample sizes to be 50, 100, 600, and 850. The first two experiments (*n*s = 50 and 100) reflect small samples and the latter two (*n*s = 600 and 850) reflect relatively large samples in a typical empirical application.[12] Figure 7 reports the results based on the above-noted Monte Carlo experiments for Model 1 to 6. Figure 7a summarizes Monte Carlo results for Models 1 to 3 in the four different experiments. In all four experiments, the model includes the full Bayesian IRT index (Model 3) outperforms the



(a) Model 1-3 (Performance Measure: The Logged Number of ER Visits)



(b) Model 4-6 (Performance Measure: The Logged Number of Outpatient Visits)

Figure 7. Comparing relative bias of parameter estimates for the three networking indices.

other two models in terms of producing the smallest relative parameter bias. In the two experiments with relatively large samples sizes (*n*s = 600 and 800), relative parameter bias introduced by Model 1 (the model with the factor index) is not very large (less than 4%), but still noticeable. The Bayesian index reduces relative parameter bias in a more substantial way when sample size is relatively small. When the sample size is 50 (i.e., very small), the factor index introduces a substantial amount of relative parameter bias (roughly about 7%), while the Bayesian index is associated with a quite small relative bias (roughly 2.2%). Figure 7b reports Monte Carlo results for Models 4 to 6. Model 6 is the one that uses the full Bayesian networking index. The comparison is quite similar to that shown in Figure 7a. Overall, with a large sample size, all three networking measures perform fine, but the full Bayesian index yields the least amount of relative parameter bias. The influence of measurement bias associated with the factor index (reflected in Model 4) is substantial with a small sample size. Consider Figures 7a and 7b together, and it is clear that the full Bayesian IRT networking index has an advantage over the factor-analytic approach in terms of reducing the influence of measurement bias in statistical models. This advantage is particularly evident when the number of observations is relatively small.

Concluding Discussion

The Bayesian IRT approach to measurement have three key features: the requirement of specifying assumptions about both the measurement items and the latent traits, statistical inferences in probabilistic (distributional) terms, and the flexibility in building complex models to explicitly acknowledge uncertainty in measurement. The requirement of specifying assumptions is particularly useful where measurement theory and conceptions of dimensionality are limited in the literature. The Bayesian IRT approach to measurement is particularly useful to deal with latent traits, which are not direct observable, or have multiple dimensions. It helps to improve measurement reliability and reduce the influence of measurement bias in statistical analysis. The unsettled nature of measurement in relation to collaborative networking is but one example. One could also find this approach useful in other areas with unsettled measurement such as red tape, public service motivation, and organizational performance.

The flexibility of the Bayesian IRT approach allows researchers to develop poorly suited models, especially when researchers operate with small samples and are uncertain about the unknown parameters. The Bayesian approach to measurement is not, though, a panacea to all measurement errors. Bayesian IRT models cannot themselves overcome the limitations of the input data. If the input data includes significant missing data or a small sample (limiting statistical power), strong skew (reducing variation and again limiting statistical power), or unreliable questions (resulting in either high variance or systematic bias in the responses), the Bayesian IRT approach will only allow one to incorporate these problems but will not produce simple or useful measures. The advantage in these cases is that Bayesian IRT makes these problems transparent rather than concealing their presence behind the assumption that resulting factors will always have a mean of zero and a standard deviation of one.

Just as any other statistical approach, researchers often face the trade-off between model flexibility and complexity when using the Bayesian methods. While there is no restriction to building complex model specifications in the Bayesian setup, computation procedures could become time-consuming and difficult as the number of parameters increases in a model. Bayesian models, moreover, cannot be implemented "in a cookbook fashion" (Gill & Witko, 2013, p.485) using off-the-shelf packaged routines in popular statistical packages such as STATA and SAS. When applying the Bayesian approach to measurement bias, researchers need to consider the specific empirical context and carefully develop model assumptions. Various computation tools, such as JAGS, WinBUGS, Stan, and many R packages, are freely available to interested researchers who want to implement Bayesian models using MCMC algorithms.[13]

Our primary interest is to provide an alternative to the factor-analytic techniques now common in public administration research. Factor-analytic techniques provide powerful tools for addressing the measurement problems present in key areas of public administration research but bring with them a number of constraining assumptions (often ignored or otherwise concealed from readers). As an alternative, we propose a Bayesian IRT approach to measurement, where these assumptions are transparent and chosen by the researcher rather than imposed by the technique. The appropriate technique will depend on which assumptions are reasonable within the measurement model. If one can reasonably assume that each component item has no error correlation with other items and has homogeneous item difficulty, the traditional factor analysis approach may be appropriate. If differential item difficulty is present, one should move to an IRT approach—a Bayesian one, in particular, if inter-item error correlation may be a problem or if the sample size is small. Nevertheless, measurement errors might vary across different empirical contexts. If the Bayesian approach is to be correctly applied to deal with measurement errors, researchers need to carefully make assumptions about measurement and the underlying data-generating process. We do not recommend the fully Bayesian alternative in all cases, but the choice of simpler measurement models should be considered (and justified) carefully.[14]

Of course, this approach is only the beginning of the use of Bayesian IRT in public administration research. We anticipate extending this analysis to including a discussion of missing data in measurement models and the intersection of imputation and IRT models. Furthermore, we would like to compare the measurement of external networking with the allocation of time connecting to internal nodes to assess the degree (or existence) of a trade-off between networking and internal management. Finally, we would like to compare various measurement approaches in the context of performance models. Does one measurement approach create measures that better explain organizational performance? What does this tell us about the nature of organizational performance and the survey reporting of such performance.

The field of public administration has not paid a great deal of attention to issues of measurement (Torenvlied et al., 2013; Whitford & Meier, 2013). As quantitative research extends into areas of study where existing, validated measures do not exist in other fields, it is important for public administration scholars to take measurement seriously. As this article illustrates, the Bayesian IRT approach provides an alternative to carefully assess measures and create what may turn out to be more effective elements for future research. Bayesian measurement applications using other substantive datasets (e.g., Bertelli, Mason, Connolly, & Gastwirth, 2013) will be useful for generalizing knowledge about how to deal with measurement issues in public administration research. As Gill and Witko (2013) prescribe, "the Bayesian approach will continue to gain in popularity" owing to the increase of Bayesian models in the social sciences and various freely available computation tools (p.461). Public administration scholars need to and will increasingly embrace the Bayesian methods as valuable tools to address methodological issues.

Authors' Note

An earlier version of this paper was presented at the 70th Annual Meeting of the Midwest Political Science Association (Chicago, Illinois), April 2012. All errors remain the responsibility of the authors.

Acknowledgment

We thank the editor and the four anonymous reviewers for their thoughtful comments.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Scholars do not all use conceptualizations of collaboration based on contact frequency. McGuire (2001) and Agranoff and McGuire (2003), for example, asked city managers to report the nature of their vertical and horizontal collaborative activities rather than the frequency. The city managers reported collaboration in information sharing, project-based partnerships, or other specific activities. Thomson, Perry, and Miller (2009) measure collaboration using attitudinal questions about collaborative experiences of stakeholders. Responses to 17 survey questions on experiences with undifferentiated partner organizations were grouped into five categories. These five categories all contribute to a single dimension of collaboration. Like the contact frequency approach, factor analysis is used to create a single dimensional measure of collaboration from a set of ordinal items—that is, attitudes related to collaboration experiences.
2. For all three models, we estimate the networking indices based on unidimensional latent scales.
3. For example, the Agency for Healthcare Research and Quality (under the Department of Health and Human Services) released the 2012 Hospital Survey on Patient Safety Culture among American Hospitals, and data were reported from 1,128 hospitals, which covers a comparable proportion of hospitals in our sample. As for all surveyed hospital staff respondents, the participation rates based on two major areas are relatively low (surgery: 10% and Medicine: 12%) (see http://www.ahrq.gov/qual/hospsurvey12/).
4. Three nodes are excluded from analysis due to extremely low response rate. These three nodes are *union representatives, volunteer board*, and *other vendors*.
5. There are different IRT models to scale polytomous survey responses. For example, one can also use the Graded Response Model (GRM) as an alternative to the Generalized Partial Credit model (GPCM) specification. In our data context, the GPCM specification fits the data better, hence we choose GPCM instead of estimating a GRM. We report detailed model comparison between the GPCM and GRM specification in the online Statistical Appendix (available at http://lingzhu2012.wordpress.com).
6. Our model produces 12 discrimination parameters (α_j) indexed by the 12 networking nodes, and 60 ($j \times (k-1) = 12 \times 5$) difficulty parameters because we constrain $\beta_{j,1}$ to be 0.
7. We choose to specify noninformative priors because they appear to be more objective than informative priors. Noninformative priors, however, still make distributional assumptions about all estimated parameters. In this empirical illustration, we follow the common practice in Bayesian IRT analysis to specify the priors for most parameters to be normal distributions (Fox, 2010; Jackman, 2009). The normality assumption is also to make sure that estimated posterior distributions for θ, α, and β are all normal distributions. As such, the results can be directly compared with those obtained using the classical factor-analytic approach.
8. See more detailed discussion on the identification and estimation of the Bayesian GPCM in the online Statistical Appendix. In areas where high-quality empirical research exists, researchers can also specify informative priors for the unknown parameters, using parameter estimations from prior empirical studies. See, for example, Jeff Gill and Christopher Witko's (2013) methodological prescription for the field of public administration.
9. Both emergency room (ER) and outpatient visits are standard performance measures to track a hospital's service outputs. The purpose of this analysis is not to include an exhaustive list of hospital performance measures. Rather, we use these two objective output measures to illustrate and compare how managerial networking is associated with organizational performance.
10. Without the log-transformation, the dependent variables are measured as count data, we estimate additional Poisson regression models to check for robustness. The results of the Poisson specification are qualitatively the same as the ordinary least square (OLS) specification.
11. Note that our empirical dataset has a large number of observations. Therefore, replacing the factor index by the two IRT indices does not drop the root mean square error (RMSE) statistics substantially. With small-*n* datasets, the changes in RMSE would be much greater.
12. We describe the specific steps for our Monte Carlo experiments in the online Statistical Appendix.

13. For an excellent introduction to Bayesian models with specific attention to Bayesian IRT measurement models see Jackman (2009).
14. For example, structural equation models (SEM) have been developed to explicitly deal with measurement bias (Fornell & Larcker, 1981). The Bayesian approach can also be applied to the class of SEM models (Lee, 2007; Scheines, Hoijtink, & Boomsma, 1999), including factor analysis (Quinn, 2004).

References

Agranoff, R., & McGuire, M. (1999). Managing in network settings. *Public Administration Review*, *16*, 18-41.

Agranoff, R., & McGuire, M. (2003). *Collaborative public management: New strategies for local government*. Washington, DC: Georgetown University Press.

Akkerman, A., & Torenvlied, R. (2011). Managing the agency environment: Effects of network ambition on agency performance. *Public Administration Review*, *13*, 159-174.

Anderson, E. B. (1997). The rating scale model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 67-84). New York, NY: Springer.

Bertelli, A., Mason, D. P., Connolly, J. M., & Gastwirth, D. A. (2013). Measuring agency attributes with attitudes across time: A method and examples using large-scale federal surveys. *Journal of Public Administration Research and Theory*. doi:10.1093/jopart/mut040

Bradlow, E. T., & Zaslavsky, A. M. (1999). A hierarchical latent variable model for ordinal data from a customer satisfaction survey with "no answer" response. *Journal of American Statistical Association*, *94*, 43-52.

Brady, H. E. (1985). The perils of survey research: Inter-personally incomparable responses. *Political Methodology*, *11*, 269-291.

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology*, *17*, 245-260.

Doty, D. H., & Glick, W. H. (1998). Common methods bias: Does common methods variance really bias results? *Organizational Research Methods*, *1*, 374-406.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Fornell, C., & Larcker, D. (1981). Evaluating structural equation models with unobserved variables and measurement error. *Journal of Marketing Research*, *18*(1), 39-50.

Fox, J.-P. (2005). Randomize item response theory models. *Journal of Educational and Behavioral Statistics*, *30*, 189-212.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer.

Fox, J.-P., & Glas, C. A. W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika*, *68*, 169-191.

Fumiko, S. (2008). Graded response model based on the logistic positive exponent family of models for dichotomous responses. *Psychometrika*, *73*, 561-578.

Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, *1*, 515-533.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). New York, NY: Chapman & Hall/CRC.

Gill, J. (2008). *Bayesian methods: A social and behavioral science approach* (2nd ed.). New York, NY: Chapman & Hall/CRC.

Gill, J., & Witko, C. (2013). Bayesian analytical methods: A methodological prescription for public administration. *Journal of Public Administration Research and Theory*, *23*, 457-494.

Graham, J. W., & Collins, N. L. (1991). Controlling correlational bias via confirmatory factor analysis of MTMM. *Multivariate Behavioral Research*, *26*, 607-629.

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in polytomous Mokken IRT model. *Applied Psychological Measurement*, *19*, 337-352.

Henry, A. D., Lubell, M., & McCoy, M. (2012). Survey-based measurement of public management and policy networks. *Journal of Policy Analysis and Management*, *31*, 433-452.

Jackman, S. (2009). *Bayesian analysis for the social science*. West Sussex, UK: John Wiley.

Johansen, M., & Zhu, L. (2014). Market competition, political constraint, and managerial practice in public, non-profit, and private American hospitals. *Journal of Public Administration Research and Theory*, *24*, 159-184.

Kim, S. (2011). Testing a revised measure of public service motivation: Reflective versus formative specification. *Journal of Public Administration Research and Theory*, *21*, 521-546.

King, G., Murry, C. J., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, *98*, 191-207.

King, G., & Wand, J. (2007). Comparing incomparable survey responses: New tools for anchoring vignettes. *Political Analysis*, *15*, 46-66.

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. West Sussex, UK: John Wiley.

Li, Y., & Baser, R. (2012). Using R and winBUGS to fit a generalized partial credit model for developing and evaluating patient-reported outcomes assessments. *Statistics in Medicine*, *31*, 2010-2026.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *23*, 157-162.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Oxford, UK: Addison-Wesley.

Masters, G. N. (1988). Analysis of partial credit scoring. *Applied Measurement in Education*, *1*, 279-297.

McGuire, M. (2001). Managing networks: Propositions on what managers do and why they do it. *Public Administration Review*, *62*, 599-609.

Meier, K. J., & O'Toole, L. J. (2001). Managerial strategies and behaviors in networks: A model with evidence from U.S. Public Education. *Journal of Public Administration Research and Theory*, *11*, 271-294.

Meier, K. J., & O'Toole, L. J. (2013). Organizational performance: Measurement theory and an application: Or, common source bias, the Achilles heel of public management research. *Journal of Public Administration Research and Theory*, *23*, 429-456.

Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research*. New York, NY: De Gruyter.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage.

O'Toole, L. J., & Meier, K. J. (1999). Modeling the impact of public management: Implications of structural context. *Journal of Public Administration Research and Theory*, *9*, 505-562.

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342-366.

Perry, J. L. (1996). Measuring public service motivation: An assessment of construct reliability and validity. *Journal of Public Administration Research and Theory*, *6*, 5-22.

Podsakoff, P. M., & Organ, D. W. (1986). Self-reports in organizational research: Problems and prospects. *Journal of Management*, *12*, 533-543.

Quinn, K. M. (2004). Bayesian factor analysis for mixed ordinal and continuous responses. *Political Analysis*, *12*, 338-353.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analysis. *Journal of Statistical Software*, *17*(5), 1-25.

Robert, C. P., & Casella, G. (2010). *Introducing Monte Carlo methods with R*. New York, NY: Springer.

Robinson, S., & Gaddis, B. S. (2012). Setting past parallel play: Survey measures of collaboration in disaster situations. *Policy Studies Journal*, *40*, 257-274.

Samijima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4, Pt. 2), 100-114.

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, *64*, 37-52.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*, 93-105.

Thomson, A. M., Perry, J. L., & Miller, T. K. (2009). Conceptualizing and measuring collaboration. *Journal of Public Administration Research and Theory*, *19*, 23-56.

Torenvlied, R., & Akkerman, A. (2012). Effects of managers' working motivation and networking activity on their reported levels of external red tape. *Journal of Public Administration Research and Theory*, *22*, 445-471.

Torenvlied, R., Akkerman, A., Meier, K. J., & O'Toole, L. J. (2013). The multiple dimensions of managerial networking. *The American Review of Public Administration*, *43*, 251-272.

Treier, S., & Jackman, S. (2008). Democracy as a latent variable. *American Journal of Political Science*, *52*, 201-217.

van der Linden, W. J., & Hambelton, R. K. (1996). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden & R. K. Hambelton (Eds.), *Handbook of modern item response theory* (pp. 1-28). New York, NY: Springer.

van Schuur, W. H. (2003). Mokken Scale Analysis: Between the Guttman Scale and parametric item response theory. *Political Analysis*, *11*, 139-163.

van Schuur, W. H. (2011). *Ordinal item response theory: Mokken scale analysis*. Thousand Oaks, CA: Sage.

Whitford, A., & Meier, K. J. (2013). Methodological innovations in public administration: A symposium. *Journal of Public Administration Research and Theory*, *23*, 307-329.

Author Biographies

Ling Zhu is an assistant professor of political science at the University of Houston. Her research interests include the political economy of welfare policy, social inequality in health care and health outcomes, the management and governance of health service networks, and research methodology.

Scott E. Robinson is the Henry Bellmon chair of public service and associate professor at the Department of Political Science at the University of Oklahoma. His research focuses on the management and politics of public agencies and the dynamics of public policy, with special attention to emergency management and administrative networking.

René Torenvlied is a professor of public management at Twente University, the Netherlands. His research interests include (formal modeling of) political-administrative processes, delegation and multilevel decision-making in democratic systems, networks and networking in public management, manifestations of red tape, as well as public performance in the fields of education, health, industrial relations, and security.