Advance Access publication June 17, 2013

*Political Analysis* (2013) 21:407–429
doi:10.1093/pan/mpt009

# An Analysis of ANES Items and Their Use in the Construction of Political Knowledge Scales

**Matthew T. Pietryka**
*Department of Political Science, University of California–Davis, One Shields Avenue, Davis, CA 95616*
*e-mail: mtpietryka@ucdavis.edu (corresponding author)*

**Randall C. MacIntosh**
*Department of Sociology, California State University–Sacramento, 6000 J Street, Sacramento, CA 95819-6005*
*e-mail: rmacintosh@csus.edu*

Edited by R. Michael Alvarez

Valid comparisons of group scores on additive measures such as political knowledge scales require that the conditional response probabilities for individuals on the observed items be invariant across groups after controlling for their overall level of the latent trait of interest. Using a multi-group confirmatory factor analysis of knowledge items drawn from American National Election Studies, we find that the scales used in recent research are not sufficiently invariant for valid comparisons across a host of theoretically important grouping variables. We demonstrate that it is possible to construct valid invariant scales using a subset of items and show the impact of invariance by comparing results from the valid and invalid scales. We provide an analysis of differential item functioning based on grouping variables commonly used in political science research to explore the utility of each item in the construction of valid knowledge scales. An application of the VTT suggests it is more appropriate to conceive of these items as effects of a latent variable rather than cause or formative indicators. These results suggest that models attempting to explain apparent knowledge gaps between subgroups have been unsuccessful because previously constructed scales were validated by fiat.

## 1 Introduction

Political knowledge is a central construct in a number of theories, including voting behavior (Palfrey and Poole 1987; Macdonald, Rabinowitz, and Listhaug 1995), political discussion (Huckfeldt 2001; Ahn, Huckfeldt, and Ryan 2010; Morehouse Mendez and Osborn 2010), persuasion (Zaller 1992), decision making (Lau and Redlawsk 2001), media consumption (Young 2004; Prior 2005), perception of racial bias (Pantoja and Segura 2003), and motivated reasoning (Lodge and Taber 2000; Taber and Lodge 2006). Research testing these theories typically assumes that survey questions measuring an awareness of officeholders and candidates can be combined with questions about party platform positions and ideological placements to create a valid scale on which sub-populations can be compared.

Recent research has sought to determine the degree to which aspects of the survey instrument limit the validity of knowledge scales (Mondak 2001; Davis and Silver 2003; Mondak and Anderson 2004; Prior and Lupia 2008; Gibson and Caldeira 2009). An unaddressed question is whether the measurement properties of such scales are sufficiently invariant across groups to justify valid comparison of sub-populations. Although consideration of measurement invariance is not new to political science (e.g., Alvarez and Nagler 2004; Clinton, Jackman, and Rivers 2004;

*Authors' note*: We thank Debra Leiter and Ron Rapoport for helpful comments and Jay Dow for sharing replication materials. Supplementary materials for this article are available on the *Political Analysis* Web site (Pietryka and Macintosh 2013).

Davidov 2009; Stegmueller 2011), we could find no instance in the literature where a rigorous test has been applied to support the assumption that political knowledge scales can be the basis for valid cross-group comparisons—despite the theoretic importance of political knowledge to the discipline.

The purpose of this study is to fill that void by assessing the extent to which recently published political knowledge scales constructed from American National Election Studies (ANES) data can be used to make valid comparisons. We rely on a well-developed literature in the social sciences on measurement invariance spanning three decades (Muthen and Christoffersson 1981; Byrne, Shavelson, and Muthen 1989; Horn and McArdle 1992; Meredith 1993; Widman and Reise 1997; Vandenberg and Lance 2000; Millsap and Yun-Tein 2004) and readily available software to estimate the necessary models (Jöreskog and Sörbom 2006; Arbuckle 2009; Muthen and Muthen 2010). We show that recently used knowledge scales are not valid for group comparisons across a host of theoretically important grouping variables. Using a subset of knowledge items, we then provide a set of new scales that demonstrate measurement invariance for most grouping variables and compare the invalid full scales and the new invariant scales, demonstrating that commonly employed knowledge scales can produce misleading estimates of differences in knowledge between subgroups. We also examine the extent to which individual knowledge items are non-invariant across groups, providing guidance for future research by helping identify the items that are most problematic for the construction of measurement invariant knowledge scales. Finally, we use a vanishing tetrad test (VTT) to assess the nature of the measurement model underlying political knowledge items. The test suggests it is more appropriate to conceive of these items as effects of a latent variable rather than cause or formative indicators.

## 2 The Centrality of Political Knowledge

Scholars agree widely that political knowledge shapes the behavior of citizens in a democracy. The surplus of information that citizens require to participate in politics coupled with a scarcity of time and effort places a premium on political expertise (Downs 1957). Political participation requires individuals to draw from a series of political cognitions (e.g., beliefs, attitudes, associations) that rest in long-term memory. For political experts, these cognitions are well organized and contain a great deal of information spanning a wide range of policy domains (Luskin 1987). Consequently, political stimuli bring relevant information automatically to active memory (Lodge and Taber 2000; Taber and Lodge 2006). For political novices, relevant information may come less quickly to mind or may be absent entirely from long-term memory. Hence, political experts behave quite differently than non-experts in a number of circumstances.

Research shows that political expertise increases the propensity to vote and the extremity of policy attitudes (Palfrey and Poole 1987). Experts' opinions are more stable and less susceptible to influence (Lodge and Taber 2000; Taber and Lodge 2006). Moreover, expertise is a self-reinforcing condition as it governs subsequent information intake through media consumption decisions (Prior 2005) and discussion patterns (Huckfeldt 2001; Morehouse Mendez and Osborn 2010).

There is considerable debate over whether encyclopedic political knowledge is necessary for meaningful participation. One camp argues that citizens can overcome a dearth of knowledge by utilizing cues and shortcuts (Lupia 1994; Freedman, Franz, and Goldstein 2004; Franz et al. 2007). Others argue that citizens lacking some baseline knowledge cannot properly apply such cues (Lau and Redlawsk 2001), which leads to suboptimal decision making (Bartels 1996). Nonetheless, knowledge scales are the most commonly employed indicator of political sophistication and political awareness (Luskin 1987; Lodge, McGraw, and Stroh 1989; Zaller 1992; Delli Carpini and Keeter 1993).

### 2.1 *Determinants of Knowledge*

A large body of research focuses on the determinants of political knowledge. Thus, scholars seek to compare the knowledge levels of various groups. Scholars have widely accepted that cognitive ability and interest increase political knowledge (Luskin 1990; Highton 2009). Research suggests that education, on the other hand, has little or no effect on political knowledge after controlling for

its determinants (Luskin 1990; Highton 2009). There is more controversy about the other potential determinants of knowledge. Luskin (1990) finds no effect of age on knowledge, whereas Lau and Redlawsk (2008) find that age increases factual information while decreasing information processing abilities. Likewise, Luskin (1990) finds no media exposure effect, whereas Prior (2005) finds that media choice has a large effect on knowledge levels.

A recurring finding in the literature is that women are less knowledgeable than men (Delli Carpini and Keeter 1996, 2000; Verba, Burns, and Schlozman 1997; Frazer and McDonald 2003). Scholars have attributed this knowledge gap, in part, to women's lower propensity to guess (Mondak and Anderson 2004; Lizotte and Sidman 2009), interviewer effects (McGlone, Aronson, and Kobryonowicz 2006), and gender-based differences on the strength of association between personal characteristics and knowledge (Dow 2009). The gap remains after controlling for these considerations, and the estimated size of the gap varies greatly between studies. The content of knowledge scales is also under scrutiny as a growing body of literature suggests that the substance of knowledge batteries is biased toward male conceptions of politics while excluding the considerations most important to many females. Recent research suggests that the knowledge gap disappears on issues of practical relevance to women, such as social services, government benefits, and female representation in government (Stolle and Gidengil 2010; Dolan 2011). If the political wants and needs of men and women differ, then relevant political considerations should be expected to vary along gender lines as well.

Thus, there are strong methodological and theoretical barriers that may inhibit the direct comparison of knowledge levels of various groups using standard knowledge batteries. The inconsistent findings for the determinants of knowledge, such as gender, may stem from this measurement barrier. Before we can make meaningful comparisons between groups on such a scale, we need to be sure that the scale is invariant between groups. In the absence of measurement invariance, observed effects are confounded with group-specific differences in the measuring devices.

### 2.2 *Measuring Knowledge*

Due to the centrality of political expertise to the discipline, measurement of knowledge has been a chief concern (Luskin 1987; Delli Carpini and Keeter 1993; Mondak 1999; Mondak 2001). Knowledge is typically assumed to be an unobserved variable that can be measured by summing an individual's correct responses to a battery of survey items. These items come in several varieties, including general-ideological placements, issue placements, and objective knowledge items. Ideological placements ask respondents to place themselves on the seven-point left–right axis and to correctly identify positions of candidates and parties on the same scale (Delli Carpini and Keeter 1993; Dow 2009). Analogous issue-specific items are also included to measure respondents' knowledge of narrower policy domains. Objective knowledge questions tend to include items asking respondents to identify jobs of various politicians and the responsibilities of various government actors (Delli Carpini and Keeter 1993; Dow 2009).

Conclusions about relative knowledge levels of various groups can be sensitive to item format. Multiple-choice items encourage guessing, which lowers item reliability (Delli Carpini and Keeter 1993, 1184). Moreover, some groups of people are more likely to guess than others (Rapoport 1979). Thus, additive scales featuring multiple-choice questions will, on average, overestimate the knowledge of people and groups prone to guessing. Researchers can avoid differential guessing by discouraging "do not know" responses (Mondak 2001).

Open-ended items feature their own problems. First, they are more difficult to code. Partially correct responses are often treated as incorrect, which prevents discriminating between partially informed and uninformed citizens (Mondak 1999). It is also difficult to discourage "do not know" responses in such formats (Mondak 2001; Gibson and Caldeira 2009). Yet, past work suggests that multiple-choice and open-ended items perform equally well in terms of item discrimination and difficulty (Delli Carpini and Keeter 1993, 1191). Thus, commonly employed knowledge scales tend to use a mix of both formats.

### 2.3 *Items as Reflective or Cause Indicators*

There are several differing conceptualizations of how political knowledge can be measured. Much of the prior research cited above is based on the conceptual view that the level of political knowledge someone possesses can be "tested" via survey questions in a manner analogous to an academic test. At the foundation of this view is the idea that political knowledge is similar to other forms of knowledge possessed by an individual. Although the content of other various forms of knowledge differs, there are fundamental similarities on how it is acquired, stored, and accessed.

From this measurement point of view, the amount of knowledge one possesses in a domain of life is an internal trait, or a latent variable, that is not directly observable (but is assumed to be sufficiently unidimensional because it is domain specific). The latent trait can be measured by presenting questions from within that domain to a respondent. These observed indicators are a means of estimating levels of the latent trait. The item responses "reflect" the level of knowledge (Fig. 1A). Higher levels of knowledge in the domain increase the likelihood of a correct answer to any particular question, although the odds of a correct answer will differ from question to question depending on their respective difficulties. It would, for example, be easier to name the President than the Speaker of the House. But we would expect someone with a greater level of political knowledge to have a higher likelihood of getting both correct compared to someone with a lower level of knowledge. This measurement model is commonly labeled an "effects" model (Bollen and Bauldry 2011) in which the observed indicators reflect the latent trait. This is the measurement model that underlies factor analysis.

There is an alternative conceptualization in the literature.[1] Although it is not often stated explicitly as such, this view treats political knowledge items as "causes" of political knowledge rather than the effects of an internal state or trait. Within this view, there are two subdivisions of measurement models. One is that the items are causes of the latent trait (cause model). The other is that items can be combined as a substitute for a latent trait (formative or composite model).

It is under this latter subdivision that indexes are constructed as a composite measure to substitute for the latent variable.[2] As noted above, the common approach is to sum a set of dummy variable responses into a single score. At the limit, a single knowledge item is sometimes used as a predictor variable but composite indexes are more commonly utilized. The other causal approach is to treat several indicators as "causes" of the latent variable rather than combining them into a composite index (Fig. 1B). This allows for an estimate of error at the latent-variable level. As we explain in the next section, validity assessments of composites are problematic. We can, however, assess validity under both the effect or cause indicator models.

Furthermore, the distinction between an effect and cause indicator model of political knowledge is important because the causal order between items and latent construct determines which measurement characteristics are appropriate for assessing validity (Bollen and Lennox 1991).

### 2.4 *Validity Assessment for Cause and Effect Indicators*

Substantial work has been done over the past two decades on methods to judge whether indicators are best considered causes or effects of the latent variable. Work has also proceeded in methods to assess the measurement validity of cause indicators. As noted above, Bollen (2011) distinguishes between causal indicator models and composite (formative) indicator models. A composite variable, such as an index that is a linear combination of items, is assumed to reflect perfectly the level of political knowledge because of two strong assumptions: measurement error is not

[1] We are grateful to an anonymous reviewer for raising this issue.

[2] A notable example of this view is Luskin's (1987) guide to measuring political sophistication. Luskin (1987) treats political sophistication as a conjunction of the size, range, and organization of an individual's political beliefs. Luskin argues that, while these three dimensions of sophistication are related, they each capture a separate—and necessary—dimension. Thus, unlike effect indicators—where the latent variable is determined by the common variance between items (MacKenzie, Podaskoff, and Jarvis 2005)—these three dimensions (and their associated indicators) each provide a unique source of variance to the underlying construct.

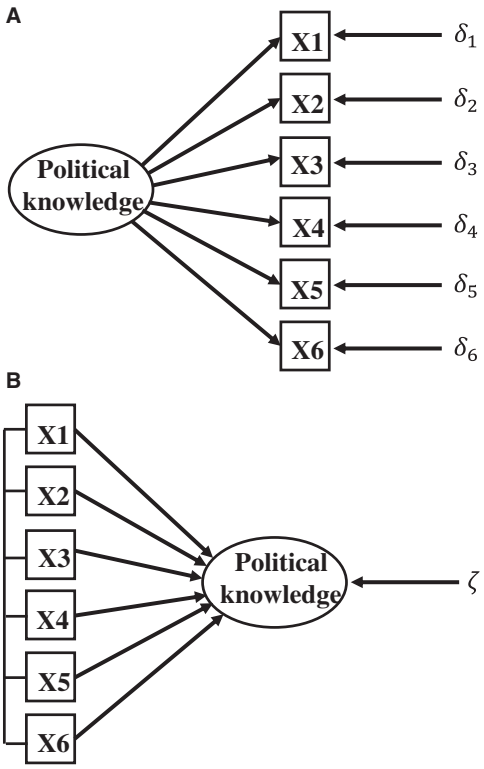


**Fig. 1** Effect and cause measurement models. (A) Six-item reflective effect model. (B) Six-item cause model.

considered for the items and the index error variance is also assumed to be zero. Indicators of composite variables have been called "formative indicators" in the past (Fornell and Bookstein 1982). We can test the assumption that the disturbance term for such an index is zero.

Bollen (2011) points out that because these indicators do not necessarily represent a unidimensional construct, it may not make sense to assess empirically their measurement validity because they are validated by fiat. He observes that if composite indicators are convenient ingredients used to form a linear composite variable, then it is not clear that assessments of validity apply to composites. Similarly, the measurement validity of a single item cannot be determined.

In contrast to a composite variable, the validity of multiple cause indicators of a latent variable can be assessed. This type of indicator also requires strong assumptions, however, about their measurement quality: they too are assumed to be measured without error at the item level. This is an untenable assumption in most situations, according to Edwards and Bagozzi (2000). Since the indicators are exogenous to the model, the error term is at the latent-variable level rather than at the item level. This latent-variable disturbance term represents all the other causes omitted from the model. A common way to validate a causal indicator is to assess the strength of its relationship to the latent variable it creates by examining the squared correlation coefficient and the structural coefficient Bollen (2011) between the item and the latent variable.

This approach is similar to what is utilized to assess the validity of effect models, for which valid comparisons require that the measurement properties of the scale are invariant across groups. It is possible to analyze the measurement properties of such items using Item Response Theory (IRT). IRT has received attention from political scientists in recent methodological texts (Jackman 2008), and it has been applied to the gender gap in knowledge, but rarely for knowledge scales more generally. Of note, Lizotte and Sidman (2009) tested 122 items from 10 ANES surveys and found that apparent observed gender differences in knowledge are confounded by the failure to obtain

cross-gender measurement invariance. They did not, however, report any combination of items that do obtain invariance. We extend that study by exploring a number of item subsets and grouping variables.

### 2.5 *Validity and Testing for Measurement Invariance*

Shadish, Cook, and Campbell (2002) define construct validity as the result of consistent operational definitions. If each operational instance (survey item) is of the same underlying construct, then the same causal relationship should result, regardless of how an outcome is operationally defined. If this result fails to be observed, this means that the operations are not in fact equivalent, and they tap into different constructs and, consequently, into different causal relationships.

For scale development with effect indicators, aspects of construct validity can be tested rigorously by assessing measurement invariance. We would contrast this approach with one that validates a scale by fiat. The definition of measurement invariance is based on the conditional probabilities of a correct answer, given the individual's common factor score (Millsap and Yun-Tein 2004). Specifically, the conditional response probabilities for individuals on the observed items should be invariant across groups after adjusting for any differences in their overall level of political knowledge, which we will refer to as the "latent trait of interest" or the common factor. In other words, there should be "local independence" for the items such that there is no *group* × *item* interaction after partialling out the trait of interest. If such an interaction exists, the item is said to exhibit "differential item functioning," or DIF, which is often labeled "item bias" in testing situations.

Ackerman (1992) describes the basis of item bias as unintended multidimensionality. A "nuisance" dimension intrudes on the measurement occasion and is responsible for the *group* × *item* interaction that is observed after partialling out the trait of interest. As this nuisance dimension is distributed unequally between subgroups, its effect introduces the item bias into the scale. The nuisance dimension can be one of a variety of factors, such as differences in item salience, prior experience, or socialization. A rigorous test of item bias is accomplished by assessing measurement invariance. Scales that are non-invariant are not sufficiently unidimensional for valid cross-group comparisons because the underlying constructs are not distributed uniformly across groups. In practical terms, a scale that is not sufficiently unidimensional cannot accurately reflect a person's performance in a single score, such as a sum or a mean value.

The model to test measurement invariance can be expressed in terms of an underlying *latent response variable* for each observed indicator, and a *latent trait* (or factor) that is of substantive interest—political knowledge in this case. The latent response variable represents the observed indicator, and the latent trait represents the overall level of knowledge as measured by the scale.[3] Each dichotomous observed indicator $X$ is related to an underlying continuous latent response variable, $X^*$, such that the item is observed to be answered correctly if that latent response variable exceeds a certain threshold, $\tau_x$.

$$X = 1, \text{ if } X^* \geq \tau_x, \text{ otherwise } X = 0. \quad (1)$$

The latent response variables ($X^*$) underlying the items are then used in a multi-group confirmatory factor analysis model to estimate the measurement and latent trait parameters.

The resulting Mplus one-factor model with a probit link is

$$P(X = 1|\eta) = \Phi\left\lfloor(-\tau_x + \lambda_x\eta)\theta^{-1/2}\right\rfloor, \quad (2)$$

where $\eta$ is the latent trait of interest (political knowledge), $\tau_x$ is the item threshold, $\lambda_x$ is the item loading, and $\theta$ is the residual variance, which is typically standardized to 1.0 (Muthen and Muthen 2010). A one-factor model is used here because prior substantive research has assumed that the political knowledge scale is sufficiently unidimensional and thus an individual's knowledge can be

[3] Subscripts for items, individuals, and groups are not shown to simplify the presentation.

summarized by a single value. The single value is either an average score or a summated score based on the number of correct answers. These values are then averaged within groups, and the observed group means are the basis for some type of regression analysis.

In contrast to the observed mean, the expected value of the underlying latent response variable $X^*$ for the reference and non-reference groups can be expressed as

$$E\left(X^*_{\text{reference}}\right) = 0 \tag{3}$$

$$E\left(X^*_{\text{non-reference}}\right) = \lambda\kappa_\eta, \tag{4}$$

where $\kappa_\eta$ is the mean of the latent trait or factor, $\eta$. The latent factor mean is fixed to zero for the reference group and freely estimated in the non-reference group. Similarly, the factor loading, $\lambda$, is assumed to be invariant across groups (Muthen and Asparouhov 2002; Brown 2006).

Measurement invariance is tested by constraining the loading ($\lambda$) and threshold parameters ($\tau$) to be equal across sub-populations. The additional model misfit created by those constraints is assessed using a robust chi-square difference test between nested models as the more restricted model is compared to a model that permits the parameters to vary between the groups.[4]

It can also be seen in the equations above why meaningful comparisons of observed means require invariance across groups for both the thresholds and loadings for the items that comprise the scale. Otherwise, observed mean differences are confounded with group differences in the measurement properties of the items used to construct the political knowledge scale (Millsap and Yun-Tein 2004). For example, the failure to obtain invariance for the factor loadings indicates the latent trait is defined differently in the various groups and they are qualitatively dissimilar (Meredith 1993). When the loadings are found to be invariant but thresholds are not, then the trait is defined consistently across groups. But it is measured using a group-specific metric, and valid comparisons cannot be made between sub-populations on either the latent response variable means or, more importantly, the observed means. That is because a necessary condition for valid comparisons on the observed means is for measurement invariance to hold on the underlying latent response variables (Millsap and Yun-Tein 2004). In other words, if an item is found to be non-invariant, a consequence is that the latent response variable is not in the same metric across the groups so that its variance should not be compared between sub-populations (Asparouhov and Muthen 2006).

It should be noted that some authors invoke "partial invariance" (Byrne, Shavelson, and Muthen 1989) and suggest that while valid comparisons cannot be made on observed means, they can be made on the latent-variable group means ($\kappa_\eta$) if "most," but not all, of the loadings and thresholds are invariant. There are several limitations of this view. First, there is no general agreement what constitutes "most." Second, it does not in any way resolve the validity problem for comparisons of the observed means or measures.

Readers familiar with IRT may note that the model expressed in equation (2) is functionally equivalent to a two-parameter normal ogive IRT model (Muthen and Asparouhov 2002). An advantage of the formulation presented here is that it allows for tests of specific measurement parameters across groups.

### 2.6 *Differentiating between Cause and Effect Indicators*

Because methods for assessing measurement validity differ depending on whether knowledge items are conceived as causes or effects, we need a way to adjudicate between the two models. It is often unclear, however, whether indicators should be best considered causes or effects of a latent variable. There are two avenues to address this question. One is the application of substantive knowledge and theory. The other is an empirical assessment of each model to see which fits the data better.

[4] Estimation is via mean and variance adjusted weighted least square (WLSMV), and the robust chi-square difference test uses the approach described in Asparouhov and Muthen (2006) based on Satorra and Bentler (1999) and Satorra (2000). See Muthen and Muthen (2010) for additional details.

Conceptual checks on whether an item should be considered a cause indicator are related to issues of face validity. The first is the extent to which the item corresponds to the theoretical definition of the latent variable. The second is whether it can be argued that a change in the item would cause a change in the latent variable, or if a change in the latent variable would cause a change in the observed indicator (Bollen and Bauldry 2011). It is clear that the items do correspond to political knowledge. It is less clear which direction the causal arrow goes. It could be argued that correct candidate placement along an ideological continuum would add to the level of political knowledge. On the other hand, it is equally plausible to argue that higher levels of political knowledge would make it more likely to be able to pinpoint a candidate's ideological position correctly. So, an empirical check might provide additional evidence on which to make a judgment.

One empirical implication for composites and cause indicators is that by definition they are required to be composed of a census of indicators. Omitting an indicator is omitting part of the construct (Bollen and Lennox 1991). So, different amalgamations of causal items create different variables that are not comparable. A political knowledge variable, for example, that includes a measure of the identity of the Speaker of the House is not the same variable as one that excludes this measure. This trait makes such measures idiosyncratic and generally not comparable across studies. This conceptualization is in contrast to the effect indicator model. If all possible questions in a domain are considered the universe of indicators of the latent trait, it is appropriate to sample from within the domain. This property emerges because equally reliable effect indicators of a unidimensional concept are essentially interchangeable. Different samples of items that cover the range of the domain are expected to obtain comparable trait-level estimates on average. By this property, political knowledge items appear to function as effects indicators: Delli Carpini and Keeter (1993, 1202) show that small subsets of knowledge items capture a large proportion of variance in scales comprised of all available items in their data set.

Moreover, recent developments in structural equation modeling provide a direct test to differentiate cause and effect indicators (Bollen and Ting 1993, 2000; Hipp and Bollen 2003; Hipp, Bauer, and Bollen 2005). The available empirical test involves pairs of covariances that are implied by two different models that incorporate the differing views of the relationship between the indicators and the latent variables. A "vanishing tetrad difference test" (VTT) (Bollen and Ting 2000) compares the two models.

Briefly, a tetrad is the difference between two pairs of item covariances implied by a measurement model involving four items (equation (5)), where $\sigma_{gh}\sigma_{ij}$ is the product of the implied covariances of items $g$ and $h$ and $i$ and $j$:

$$\tau_{ghij} = \sigma_{gh}\sigma_{ij} - \sigma_{gi}\sigma_{hj}. \quad (5)$$

Because the covariances from an effects model (where the observed indicators depend on the latent variable) can be expressed completely in terms of factor loadings and the latent-variable variance, these tetrads have an expected value of zero. In other words, the tetrads are expected to disappear in an effects model ($\tau_{ghij} = 0$). In contrast, the tetrads from a cause indicator model (formulated as a multiple indicators multiple causes (MIMIC) model[5] [Hauser and Goldberger 1971]; Fig. 2) are not expected to vanish by equaling zero because the covariances across items do not cancel out.

A tetrad model difference test statistic, distributed as a chi-square variate, is available to test the hypothesis that the cause indicator MIMIC model significantly fits the data better than the effect indicator model. If the test produces a non-significant chi square, then the model with the most vanishing tetrads (the effects) is preferred. A significant chi square suggests otherwise as the less restricted (cause) model has a better fit. This test has been implemented in a SAS macro (Hipp, Bauer, and Bollen 2005) that compares the tetrads in the two model implied covariance matrices.

We can use this method to decide between the cause and effect models of political knowledge, in order to inform our subsequent validity tests. First, though, we detail the data we will use.

[5] The cause model in Fig. 1B is not identified, so estimates cannot be obtained. The MIMIC model is identified by treating at least two of the indicators as effects of the latent variable and the remainder as cause indicators.

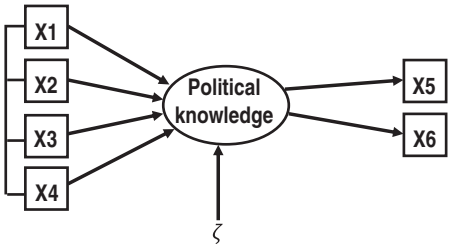


**Fig. 2** Multiple indicator multiple causes model.

## 3 Data and Measures

For this analysis, we rely on Dow's (2009) knowledge battery, which includes respondents to the 1992, 1996, 2000, and 2004 ANES.[6] The battery is largely representative of the knowledge scales commonly employed in the discipline and includes general-ideology placements, policy-specific placements, and open-ended objective knowledge questions.

For each year, there are three general-ideology items. The first equals one for the respondents who are able to place themselves on the seven-point liberal–conservative scale. The second is correct for those that place the Democratic presidential candidate at a more liberal position than the Republican candidate on the seven-point scale. The third is correct for respondents that place the Democratic Party at a more liberal position than the Republican Party.[7]

There are seven policy-specific placements for the 1992, 1996, and 2000 studies and eight policy-specific placements for the 2004 study. The policy-specific placements focus on four issues: (1) services and spending; (2) defense spending; (3) jobs and standard of living; and (4) abortion. The ANES asks respondents to identify the Democratic and Republican presidential candidates' position on each issue. The services and spending, defense spending, and jobs and standard-of-living items use seven-point scales, while the abortion items use four-point scales.[8] The 2000 study is an exception, where we collapse to a five-point scale all responses for the services and spending, defense spending, and jobs and standard-of-living items. Respondents are correct on these items if they place the Democratic candidate at a more liberal position than the Republican candidate. Respondents also place the two parties on each of these issues, though the 1992, 1996, and 2000 studies each omit one set of party-issue placements. In 1992 and 2000, the ANES does not include the abortion party placements, and in 1996 it does not include the jobs and standard-of-living party placement. Hence, the 1992, 1996, and 2000 studies have seven policy-specific items whereas the 2004 study, which includes all four party placements, has eight policy-specific items.

There are four objective knowledge items for each study year, though the content of the items varies from year to year. Each open-ended item asks respondents to identify the political position held by a public figure. For example, all four years ask respondents to identify the position held by William Rehnquist. Those answering "Chief Justice of the Supreme Court" are correct. In addition to the Chief Justice item, 1992 and 1996 studies include items asking respondents to identify the Vice President, the President of Russia, and the Speaker of the House. The 2000 study includes the Chief Justice item, as well as items regarding the British Prime Minister, the Senate Majority Leader, and the U.S. Attorney General. The 2004 study includes items regarding the Chief Justice, Vice President, British PM, and Speaker of the House.

[6] This analysis draws data from the 1992 ANES (Miller et al. 1999), 1996 ANES, 2000 ANES, 2004 ANES, and the ANES Time Series Cumulative data (The American National Election Studies 2010).

[7] For general-ideology and policy-specific placements, the ANES only asks the candidate and party if the respondent is able to place themselves on the relevant scale. Therefore, failure to self-place on any scale necessitates incorrect responses on the candidate and party placements. Similarly, failure to place either candidate results in an incorrect response for the candidate placement and failure to place either party results in an incorrect response for the party placement.

[8] The 2000 ANES uses a five-point scale for the services and spending, defense spending, and jobs and standard-of-living items for telephone interviews and a seven-point scale for face-to-face interviews. Dow (2009) translates all responses to a five-point scale; we follow his procedure.

**Table 1** ANES items used to create political knowledge scale: 1992–2004. (Adapted from Dow 2009)

| | *1992* | *1996* | *2000* | *2004* |
|---|---|---|---|---|
| Ideological | | | | |
| Self-placement | X | X | X | X |
| Candidate placement | X* | X* | X* | X* |
| Party placement | X | X | X | X |
| Objective knowledge | | | | |
| Vice president | X | X | | X* |
| Chief justice supreme court | X* | X* | X | X |
| Russian President | X | X* | | |
| British PM | | | X* | X |
| House of representatives speaker | X* | X | | X* |
| Senate majority leader | | | X* | |
| Attorney general | | | X | |
| Issue placement | | | | |
| Services and spending—candidate | X* | X* | X* | X |
| Services and spending—party | X | X | X | X* |
| Defense spending—candidate | X | X* | X* | X* |
| Defense spending—party | X* | X | X | X |
| Jobs and Std of living—candidate | X | X* | X | X |
| Jobs and Std of living—party | X | | X | X |
| Abortion—candidate | X* | X | X* | X* |
| Abortion—party | | X | | X |

* Used for six-item scale

The principal ANES investigators documented problems with the administration of some objective knowledge questions in recent surveys, particularly items that identify the Chief Justice and British PM (Krosnick et al. 2008). In light of this report, we also test whether the inclusion of these questionable items interferes with valid scale construction.

To create the scales, we sum the number of correct responses for a respondent and divide by the total number of knowledge items from the study year. The denominator in the 1992, 1996, and 2000 studies is fourteen and the denominator in the 2004 study is fifteen for the full battery.[9] (The item subsets used for each survey year are shown in Table 1.) Hence, we code respondents as incorrect on an item if they refuse to answer or provide a "do not know" response.[10] The ANES includes a pre- and post-election wave, and we draw knowledge items from both waves. Thus, we drop those respondents who did not take part in both waves. Following Dow (2009), we restrict our sample to the ANES respondents who were not missing on any of Dow's explanatory variables and omit African Americans from the analysis.

We evaluate the validity of knowledge scales for group comparisons using grouping items commonly included in models of political knowledge, including gender (e.g., Delli Carpini and Keeter 1996, 2000; Verba, Burns, and Schlozman 1997; Frazer and McDonald 2003; Dow 2009) as well as age (e.g., Luskin 1990; Lau and Redlawsk 2008), education (e.g., Luskin 1990; Highton 2009), income (e.g., Lambert et al. 1988), media use (e.g., Luskin 1990; Prior 2005), and political participation (e.g., Palfrey and Poole 1987). We compare age and income groups below the median value in a given year to those at or above the median. For education, we compare respondents with a college degree to those without a degree. We include four media-use grouping variables. The first

[9] In all years, we draw the objective knowledge items from the post-election wave. In 1992 and 1996, we draw all general-ideology and policy-specific placement items from the pre-election wave. In 2000, we draw all policy-specific placement items from the pre-election wave and all general-ideology placement items from the post-election wave. In 2004, we draw all general and policy placement items from the pre-election wave, with the exception of abortion placements, which the ANES asked only in the post-election wave.

[10] Treating "do not know" responses as missing follows convention in the literature (see, e.g., Delli Carpini and Keeter 1993; Prior 2005; Dow 2009), but, as discussed above, may inflate differences between groups due to differential propensities to guess (Mondak 2001).

three are binary measures indicating whether the respondent obtained campaign news from newspapers, radio, or television. The fourth asks the number of days in the last week the respondent watched TV news and separates the respondents below the median from those at or above the median. The participation variable separates those who participated in a recent campaign in some way (tried to influence others, attended meetings/rallies, worked for party/candidate, displayed sticker/button, donated money) from those who did not.

### 3.1 *Analytic Plan*

The analytic plan is to assess the measurement invariance of a political knowledge scale used in prior research and is based on the effect indicator model. If the full scale fails to obtain invariance across groups, we will search for a subset of items that will provide a more valid scale. Finally, we will check our model specification by conducting the tetrad test to determine whether it may be more appropriate to consider the invariant items as cause indicators. Part of the process necessary to carry out the tetrad test is the estimation of MIMIC models with these items. A useful byproduct of those models is a parameter estimate of the latent-variable error term. If that disturbance term is significant, then the formative composite index assumption of errorless measurement is ruled out (Bollen and Bauldry 2011).

## 4 Results

### 4.1 *A Test of Measurement Invariance*

To assess measurement invariance, we first estimate the non-invariance model, which is unidimensional and allows the loading and thresholds to vary by demographic group. The test of invariance is based on the degradation in model fit when the two sets of measurement parameters are constrained to be equal across groups.[11] Statistically significant differences in model fit suggest that measurement invariance fails as the misfit is large relative to the number of cross-group parameter constraints imposed. The results of tests of measurement invariance across the grouping variables are shown in Fig. 3. Table A1 provides the results in tabular form.

The figure plots the *p*-value associated with the chi-square test of change in model fit for the full and six-item scales by each grouping variable in each year. Points with values $> .05$ indicate that a scale is invariant with respect to the associated grouping variable, whereas points below this threshold indicate a scale's measurement noninvariance and therefore its invalidity for comparisons across subgroups created by that grouping variable. The figure shows that measurement invariance is rejected for scales comprised of all items for almost all demographic grouping variables as the constraints lead to statistically significant degradation in model fit. The exceptions are participation in 1996, radio use in 2004, and television use in all years. As mentioned above, the ANES principal investigators noted coding errors in several objective knowledge questions. With one exception, excluding these questionable items had no effect on the substantive results, as the results from the thirteen-item scales in Fig. 3 display. The lone exception is frequency of television viewership in 2004, in which the thirteen-item scale is invariant whereas the fifteen-item scale is not.

Figure 3 demonstrates that the full scales are not valid for group comparisons on most group indicators. We therefore also construct a new set of scales, and in the figure, demonstrate their validity for group comparisons within each year. Table 1 lists the specific items used for each year, which were selected because together they produce scales that are invariant for many grouping variables while keeping the item content similar across years. Each scale includes two objective knowledge items, three issue placements, and one general-ideological placement of the candidates. The six-item scales thus retain the diversity of items found in the full scales and the relative use of objective knowledge, issue placements, and general-ideology placements are roughly proportional between the six-item and full scales. Furthermore, there is a core subset of items that is asked in

[11] This is a likelihood ratio statistic, which is distributed as a chi-square variate with degrees of freedom equal to the number of constrained parameters.

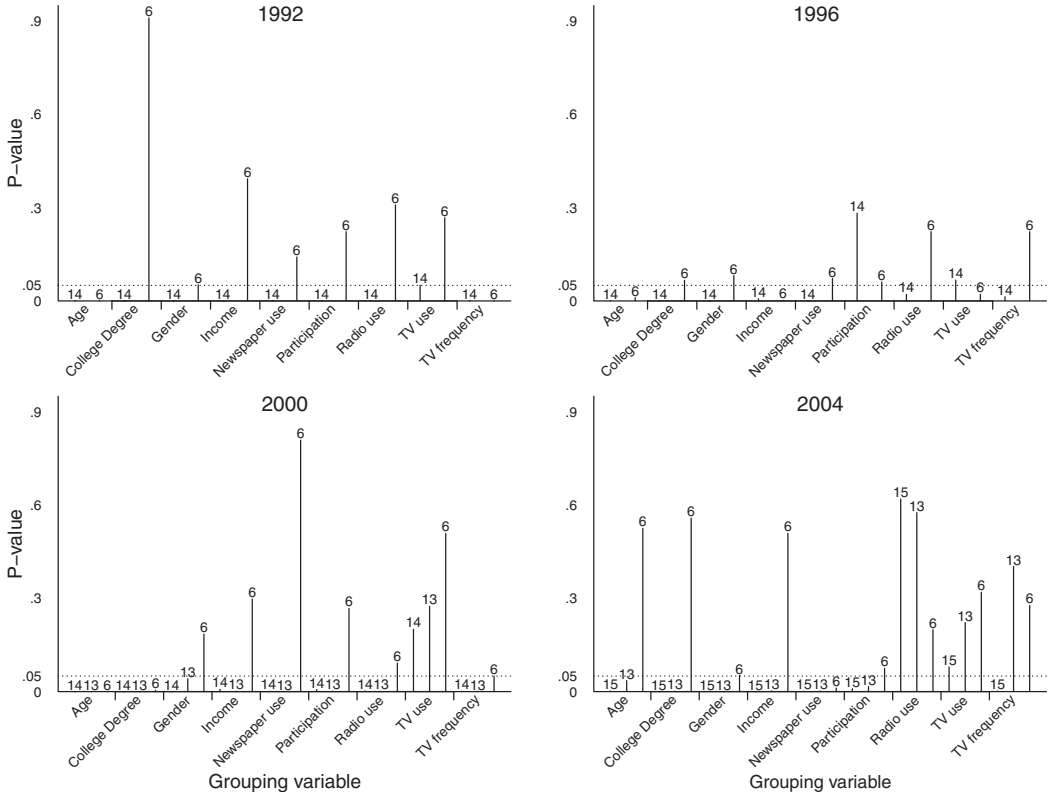


**Fig. 3** *p*-values associated with robust chi-square difference test of change in model fit when the two sets of measurement parameters are constrained to be equal across groups. Scales with *p*-values $> .05$ are said to be invariant with respect to the associated grouping variable, whereas scales falling below this threshold are non-invariant. Numbered points indicate the number of items included in each scale. The full scales include fourteen items in 1992–2000 and fifteen items in 2004. The new scales include six items in all years. The thirteen item scales in 2000–2004 remove from the full scale items with administration problems documented by the ANES (Krosnick et al. 2008). *Source*: Table A1.

every survey year. Nonetheless, we could not include all the same items in each year's scale, due to the inconsistencies in ANES items across years[12] and the ANES administration problems detailed by Krosnick et al. (2008). Yet, as we will discuss below, the inclusion or omission of any particular item would not change the substantive meaning of the scale, under the assumptions of an effect measurement model. For each year, the new six-item scales are sufficiently invariant on the basis of respondent gender, education, income, newspaper use, participation, radio use, and TV use, with a few year-specific exceptions.[13] The six-item scales are similarly invariant for frequency of TV use in 2004 and 1996, but not 2000 or 1992.

Table 2 shows that the new scales perform as well as the full scale in criterion tests of construct validity for the various ANES indicators of political participation, which constitute some of the most widely studied effects of political knowledge (e.g., Verba, Schlozman, and Brady 1995; Deli

[12] Several inconsistencies in the ANES administration of knowledge items from year to year make comparisons between years problematic. Chief among those is the five-point scale that the 2000 study uses on the services and spending, defense spending, and jobs and standard-of-living items. Due to the reduced number of positions along the scale (five instead of seven), a greater number of ties occur where respondents place the Democratic and Republican candidates (or parties) at the same position on the scale. In such instances, respondents do not receive credit for a correct answer, reducing the mean score for the year. Similarly, the point in the study (pre- or post-election) that the ANES administered various questions changes from year to year. Thus, campaign effects may create differences in response patterns for such items.

[13] The exceptions are income in 1996, education in 2000, and newspaper use in 2004.

**Table 2** Point-biserial correlations between knowledge scales and participation indicators

| | *1992* | | *1996* | | *2000* | | *2004* | |
|---|---|---|---|---|---|---|---|---|
| | *Full scale* | *Six items* | *Full scale* | *Six items* | *Full scale* | *Six items* | *Full scale* | *Six items* |
| Did R report voting? | .37 | .34 | .36 | .31 | .37 | .35 | .41 | .39 |
| Did R try to persuade anyone to vote for or against a candidate? | .13 | .13 | .27 | .25 | .24 | .24 | .25 | .26 |
| Did R wear a campaign button, etc.? | .16 | .14 | .13 | .14 | .17 | .17 | .19 | .18 |
| Did R go to any political meetings, etc.? | .14 | .14 | .14 | .12 | .14 | .14 | .13 | .13 |
| Did R do any work for one of the candidates? | .08 | .08 | .12 | .11 | .08 | .07 | .11 | .10 |
| Did R give money to an individual candidate? | .18 | .18 | .15 | .16 | .19 | .19 | .24 | .22 |
| Did R give money to a political party? | .14 | .13 | .16 | .16 | .18 | .15 | .21 | .22 |
| Did R give money to any other political group? | .20 | .21 | .19 | .17 | .16 | .17 | .19 | .18 |

Carpini and Keeter 1996; Popkin and Dimock 1999; for a review of this literature, see Galston 2001). The table shows that the point-biserial correlations between (1) the full scales and the individual participation indicators; and (2) the six-item scales and the individual indicators of participation are similar in magnitude and direction. Thus, the table suggests that these abbreviated scales can be useful for studying political knowledge, while avoiding the problem of measurement non-invariance. Moreover, the non-invariance findings suggest that the correlations between the full scales and the participation indicators are biased by some nuisance dimension. Thus, the similarity in correlations belies an important distinction between the two sets of scales. We therefore turn next to assessing the extent and direction of the bias.

### 4.2 *Measurement Non-Invariance and Effect Size*

The use of these commonly employed, but non-invariant, knowledge scales can bias considerably the estimated differences between subgroups. To evaluate the effect that measurement invariance has on substantive findings, we compare bivariate effect sizes of various factors on political knowledge estimates. Table 3 shows how measurement non-invariance confounds estimates of effect size. The table shows the effect size estimates (Cohen's *d*) for group comparisons on the basis of dichotomous measures of gender, income, political participation, and educational attainment. Here, the effect size inflation ranges from a low of 2% to as much as 13%. But there is also sizeable estimation *deflation* as it attenuates the effect sizes for gender and age in some years that range from a low of −13% to a maximum of −175%. These discrepancies are due to *group* × *item* interactions that are related to some unidentified nuisance dimension that interferes with the measurement instrument and artificially inflates or deflates the effect sizes.

This analysis demonstrates that the influence measurement non-invariance has on the estimates of effect sizes is often substantial, but also varies widely depending on the year and the grouping variable under consideration. The important point is that there is no way to predict how effect size will be confounded by non-invariant measures in any particular situation and, therefore, no way to anticipate its direction or adjust for it.

Table 4 provides another way to evaluate the consequences of non-invariance, by regressing the full scale on each of the grouping variables first without controlling for the invariant six-item scale and then after including the control. If the non-invariant scale and the invariant scale we propose are essentially equivalent, then there should be no significant coefficients in the second model. If there are, however, significant coefficients in the second model for a given year, it indicates that the full scale is contaminated by a nuisance dimension associated with the grouping variable, but independent of political knowledge—thereby biasing full-scale estimates of political knowledge differences. For instance, the statistically significant coefficients on gender in the second models in 1992 and 2000 are a third the size of the estimates from the first models for each year. Therefore,

**Table 3** How measurement non-invariance influences bivariate effect sizes

| | *Non-invarient full scale* | *Invarient six items* | *Effect size Inflation (%)* |
|---|---|---|---|
| | *Cohen's d* | | |
| 2004 | | | |
| Gender | .32 | .58 | −84 |
| Income | .58 | .52 | 11 |
| Participation | .56 | .57 | −2 |
| Education | .94 | .90 | 4 |
| Age | .06 | .18 | −175 |
| 2000 | | | |
| Gender | .55 | .48 | 13 |
| Income | .74 | .65 | 12 |
| Participation | .61 | .60 | 2 |
| 1996 | | | |
| Gender | .35 | .39 | −13 |
| Education | .95 | .85 | 10 |
| 1992 | | | |
| Gender | .45 | .39 | 13 |
| Income | .66 | .60 | 8 |
| Participation | .61 | .59 | 4 |
| Education | 1.16 | 1.10 | 5 |

*Note*. The effect size inflation measures the extent to which the non-invariant full scales inflate (or deflate) the effect size as measured by the invariant six-item scales. Aside from media-use indicators (not shown here), the full scale is not invariant for any grouping variables except participation in 1996. The six-item scales are invariant for all grouping variables except income in 1996, education in 2000, and age in 1992–2000.

the table suggests that roughly a third of the estimated gender gap in those years is due to measurement contamination. Likewise, in 1992, this contamination biases upward the estimates associated with education, participation, and radio use. Similar problems arise for the other grouping variables in many years. As with the bivariate analysis, we find the nuisance dimension affects the regression coefficients and substantive conclusions in unpredictable ways. Again, the only way to overcome this threat to validity is to use invariant measures.

### 4.3 *Differential Item Functioning*

Although the six-item scale is sufficiently invariant for many group comparisons, researchers constructing their own scales are likely to find it valuable to understand which individual items are most useful—and which are most problematic—for group comparisons. To that end, Fig. 4 shows the absolute DIF of each item, averaged across the four years under study.[14] DIF is typically measured as the difference in the item difficulty based on group membership. These IRT difficulty estimates are computed from the confirmatory factor analysis coefficients obtained from a metric invariance model for each year's full scales. The factor loadings are constrained to be equal across groups, but thresholds are group specific. The first panel of the figure averages the absolute DIF coefficient values across all grouping variables, whereas the other two panels show separately the average DIF for comparisons by education and gender.[15] As discussed above, the DIF indicates the magnitude of the item bias due to a *group* × *item* interaction that remains after partialling out the trait of interest. Thus, larger DIF values indicate items are more problematic for measuring differences in political knowledge across subgroups because of the intrusion of nuisance dimensions. Conversely, smaller DIFs indicate the items that may best combine to produce invariant scales.

[14] Table A2 of the supplementary appendix provides further information about these estimates.
[15] Figure A1 in the supplementary appendix provides the plots for all grouping variables.

**Table 4** OLS regressions of full scale on grouping variables and invariant six-item scales

| | *1992* | | *1996* | | *2000* | | *2004* | |
|---|---|---|---|---|---|---|---|---|
| | *Model 1* | *Model 2* | *Model 1* | *Model 2* | *Model 1* | *Model 2* | *Model 1* | *Model 2* |
| Six-item scales | | 0.81* | | 0.83* | | 0.87* | | 0.97* |
| | | (0.012) | | (0.013) | | (0.014) | | (0.019) |
| College degree | 0.18* | 0.02* | 0.18* | 0.04* | | | 0.15* | 0.01 |
| | (0.016) | (0.007) | (0.015) | (0.007) | | | (0.020) | (0.009) |
| Gender | −0.06* | −0.02* | −0.03 | 0.01 | −0.06* | −0.02* | −0.04* | 0.01 |
| | (0.014) | (0.006) | (0.014) | (0.006) | (0.017) | (0.007) | (0.018) | (0.008) |
| Income | 0.05* | 0.01 | | | 0.11* | 0.02* | 0.03 | 0.02* |
| | (0.015) | (0.007) | | | (0.017) | (0.008) | (0.019) | (0.008) |
| Newspaper use | 0.10* | 0.01 | 0.08* | 0.02* | 0.10* | 0.01 | 0.07* | 0.00 |
| | (0.015) | (0.007) | (0.015) | (0.007) | (0.018) | (0.008) | (0.020) | (0.009) |
| Participation | 0.09* | 0.02* | 0.07* | 0.01 | 0.09* | 0.01 | 0.06* | 0.01 |
| | (0.014) | (0.006) | (0.015) | (0.007) | (0.017) | (0.007) | (0.018) | (0.008) |
| Radio use | 0.06* | 0.01* | 0.07* | 0.01 | 0.12* | 0.01 | 0.08* | 0.02* |
| | (0.014) | (0.006) | (0.015) | (0.007) | (0.017) | (0.008) | (0.018) | (0.008) |
| TV use | 0.05 | 0.00 | | | 0.05* | 0.00 | 0.07* | −0.01 |
| | (0.024) | (0.011) | | | (0.023) | (0.010) | (0.028) | (0.012) |
| TV frequency | | | 0.01 | 0.00 | 0.04* | 0.01 | 0.01 | −0.01 |
| | | | (0.015) | (0.006) | (0.017) | (0.007) | (0.019) | (0.008) |
| Age | | | | | | | 0.00 | −0.02* |
| | | | | | | | (0.018) | (0.008) |
| Intercept | 0.38* | 0.17* | 0.50* | 0.13* | 0.28 | 0.11* | 0.46* | 0.05* |
| | (0.026) | (0.012) | (0.017) | (0.010) | (0.025) | (0.011) | (0.031) | (0.016) |
| $N$ | 1028 | 1028 | 973 | 973 | 886 | 886 | 621 | 621 |
| $R^2$ | 0.35 | 0.87 | 0.25 | 0.85 | 0.28 | 0.86 | 0.28 | 0.86 |

*Note*. *$p < .05$. OLS regression. Standard errors in parentheses. Table omits grouping variables for which the full scale is invariant. Model 2 coefficients indicate the magnitude and direction of the biases in effect sizes introduced by the non-invariant scales.

A general conclusion from the figure is that many of the items commonly included in knowledge scales are problematic. In particular, open-ended objective knowledge items (shaded black in the figure) demonstrate some of the largest DIF values for every grouping variable. This result is not limited to the Chief Justice and British PM items—those items whose coding schemes were highlighted as most problematic by the ANES administrators (Krosnick et al. 2008)—but rather extends to all the open-ended objective knowledge items. This result provides insight into the debate regarding item format; researchers should avoid these items, instead focusing on multiple-choice formats.

Yet, the remaining items are also problematic for many of the grouping variables, with noticeable variation in DIF across groups. For instance, the small DIF values suggest that abortion items are sufficiently useful for knowledge comparisons between genders, but these items exhibit considerable DIF for comparisons between education levels. The services and spending items also demonstrate relatively low DIF values for gender comparisons, reinforcing the results of Dolan (2011) and Stolle and Gidengil (2010) and highlighting the need for researchers to think critically about which items are most relevant for the groups under comparison.

### 4.4 *Vanishing Tetrad Test*

We assume an effects measurement model in the above tests, and we therefore must test that assumption using a VTT. To carry out the VTT, a MIMIC model for each year has to be estimated. We used the two items that represent the highest level of abstraction as the reflective measures and

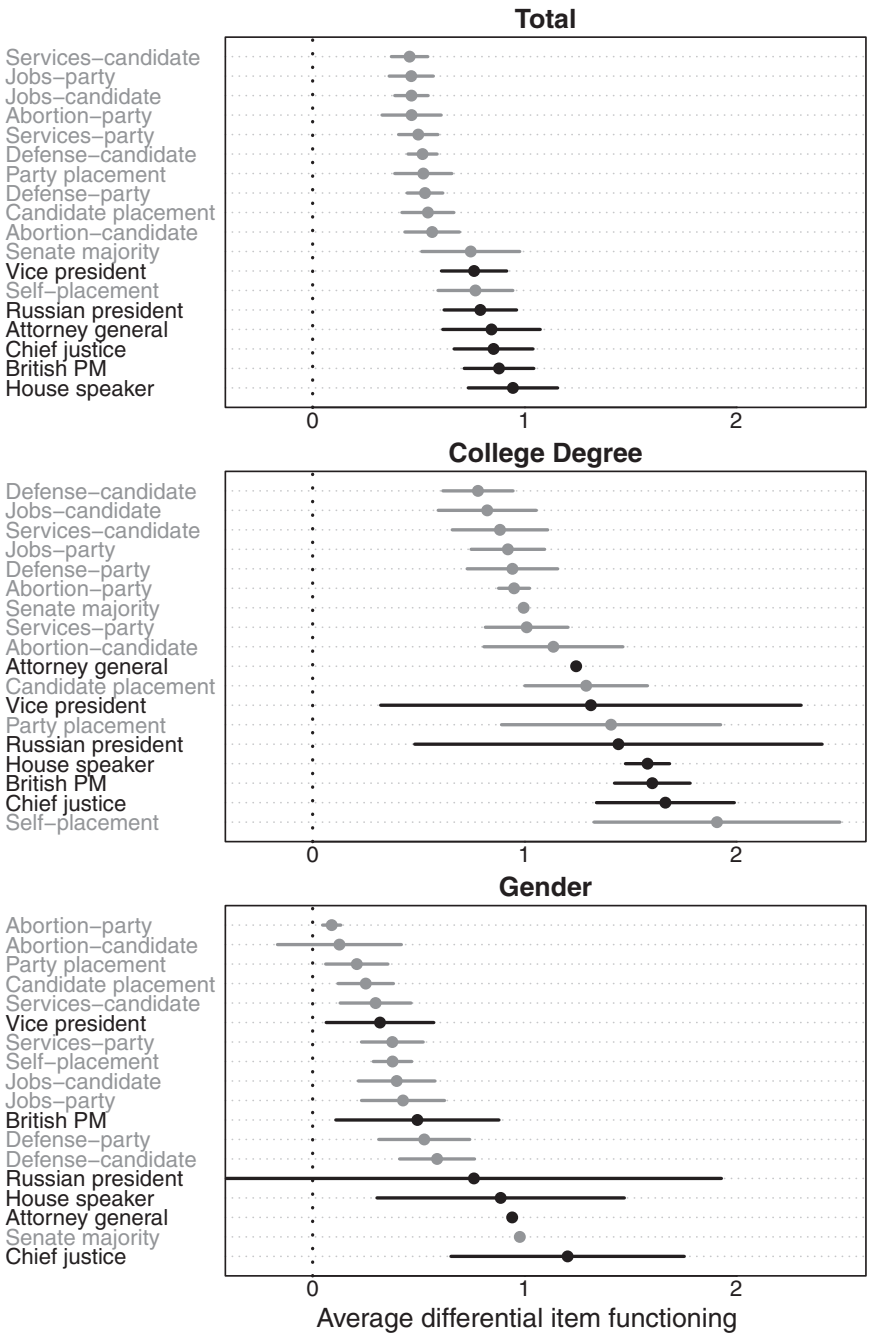


**Fig. 4** Average DIF for selected grouping variables, 1992–2004. Open-ended objective knowledge items are shaded black. *Total* panel averages across all grouping variables and all years. See Fig. A1 for plots of each grouping variable. *Source*: Table A2.

the items measuring objective facts as the causes. As such, we assume objective facts are the basis of schema that provide a more complex context for an abstract understanding of candidate, ideological, and party positions (Axelrod 1973; Conover and Feldman 1984).[16] To start, we can test

[16] It is beyond the scope of this study to review learning theories. But we cite schematic theory as one applicable example that would be familiar to many political scientists. Under this theory, constructive errors are the product of an elaboration process occurring during or shortly after encoding (facts) as general terms are stored as specific instantiations (Alba and Hasher 1983). It is the integration and interpretation of facts that leads to schematic formation and linkages.

**Table 5** VTT chi-square values for 1992–2000[a]

| | *Effect model* $\chi^2$ | *df* | *MIMIC model* $\chi^2$ | *df* | *VTT* $\chi^2$ | *df* | *p-value* |
|---|---|---|---|---|---|---|---|
| 1992 | 6.1 | 8 | 1.6 | 3 | 4.5 | 5 | .48 |
| 1996 | 8.8 | 8 | .4 | 2 | 8.4 | 6 | .21 |
| 2000 | 9.4 | 7 | 6.4 | 3 | 2.9 | 4 | .57 |

*Note*. A statistically significant VTT $\chi^2$ indicates that the cause model best fits the data. A non-significant $\chi^2$ fails to reject the effect model in favor of the cause model.
[a]The best-fitting models for 2004 were not tetrad nested so the VTT could not be carried out for that year.

whether the formative composite assumption that these subsets of invariant items would constitute an index with no measurement error is supported by the data. We find that the latent-variable error terms for the MIMIC models (not shown) are all significant. These results rule out the composite index model as it is inconsistent with the data.

Recall that the cause (MIMIC) model is tetrad nested within the effect model. All the tetrads are expected to vanish in the effect model, and none of the tetrads are expected to vanish in the cause model. Since the cause model is nested within the effect model, a significant chi-square statistic, based on the discrepancies between what is observed and what is expected, lends support for the cause model as fitting the data better. A non-significant chi square fails to reject the effect model in favor of the cause model.

Table 5 shows that for three survey years, the effect model is the preferred model.[17] All of the nested VTT chi-square values, which are the differences between the effect and cause model values (shown at the right of Table 5), are not significant. We conclude from these tests that it is more appropriate to treat the invariant subset of knowledge items as a reflection of the latent variable rather than the cause.

## 5 Discussion

Given the centrality of political knowledge, we are surprised by the lack of utility that these measures provide for key questions in the field. The results presented here call into question a fundamental assumption that underlies a substantial body of research: that combining knowledge items constitutes a measuring device that permits valid group comparisons.

Focusing on the gender analysis, it appears that qualitative differences in how women and men define and pursue political information, coupled with the differential survey response patterns discussed above, present a challenge in the creation of an invariant knowledge scale. Women seem to have different political preferences and priorities than do men (e.g., Smiley 1999; Stolle and Gidengil 2010), and each gender consumes political information in vastly different ways (e.g., Yum and Kendall 1988; Huckfeldt and Sprague 1995). All of these distinctions condition the relevance of particular types of political information. Such differences may impede the construction of invariant knowledge scales because they can cause individual scale items to be differentially linked to gender, as we see here.

Turning to knowledge comparisons for other grouping variables, the results presented here provide six-item subsets for valid comparisons on the basis of education and participation. In

The subsequent integration of a greater number of schema leads to more abstract constructions and generalizations and enhances the ability to retrieve the facts underlying the schema. There are other learning theories based on developmental stages that lead to increasing complexity and hierarchical retrieval system models, as well. All of the aforementioned learning theories, however, are based on the idea that synthesis reflects a more complex internal process than the retrieval of a single fact.

[17] For the years 1994–2000, we could utilize effect models for the vanishing tetrad test that specified a correlated error between two effect indicators. The 2004 data, however, require two such correlated error parameters to obtain an acceptable model fit. Unfortunately, that effect model is not tetrad nested within the MIMIC model and prevents its application of the VTT. The effect model is preferred over the causal indicator MIMIC model in the three years that we could test.

addition, these scales also provide possibilities for valid comparisons by income, with the exception for 1996.

We were able to find satisfactory subsets for comparisons on the basis of newspaper use for all years except 2004. Similarly, comparisons on the basis of TV use (excepting 1996) and radio use appear to be valid using six-item measures. We note that in contrast to what is observed for other demographic variables, the full scale is invariant in several years for radio and television use. This result leads us to question the causal ordering with regard to the electronic media-use variables given what we observe for other demographics. A plausible argument could be made that political knowledge leads to more (or different) media usage, rather than media usage leading to enhanced political knowledge. If media use and political knowledge are the joint effects of other factors, such as previous political participation, then the media-use variables cannot be the source of item bias in the knowledge scale because they are concurrent or subsequent in the causal chain. This expectation is consistent with our results and suggests an opportunity for future research.

The VTT analysis also provides insight into the relationship between political knowledge as a construct and its associated measures. While the discipline has rarely stated the measurement model explicitly, Delli Carpini and Keeter's (1993, 1996) foundational work on political knowledge measurement uses classical test theory and IRT to assess validity—implicitly assuming an effect model. Our analysis tests the measurement model and provides empirical support for their assumption and lends guidance to future researchers regarding the appropriate methods for validating political knowledge scales.

This analysis calls attention to an ongoing debate in the field of structural equation modeling regarding the applicability and value of causal and formative measurement models. As scholars in a diverse set of fields have begun to recognize the importance of understanding the link between constructs and their measures, the theory underpinning cause models has received significant attention (e.g., Blalock 1964; MacCallum and Browne 1993; Diamantopoulos and Winklhofer 2001). More recently, some scholars have called the utility of cause models into question (e.g., Borsboom 2006; Iacobucci 2010; Bagozzi 2011). For example, Edwards (2011) argues that causal models rest on the often-untenable assumption that there are no errors in the measurement of formative indicators. Moreover, such models are often difficult or impossible to validate because their estimation requires researchers to make arbitrary decisions that can have important consequences for interpreting the latent construct and, likewise, the construct lacks meaning absent its heterogeneous indicators, preventing tests of construct validity. Further, Edwards (2001) argues that there are philosophical and ontological advantages to conceptualizing measures as effects and taking a critical realist position rather than taking an instrumentalist or operationalist orientation. Classical and modern test theory, along with IRT, are all based on the latent trait effects model. Among the advantages we would add to his list is the ready availability of an extensive literature of learning theory from cognitive and education psychology to explain the acquisition, retention, and activation of knowledge. Political science has utilized this literature for almost half a century in the study of political knowledge (Seeman 1966). Underlying this approach is an assumption of a latent trait that drives item responses as effects that are measured with "tests." In contrast, because cause indicators are exogenous, there is no such theory or explanation of their genesis. Another substantial advantage is that of item selection with an effect model: the application of a domain sampling model.

### 5.1 *Domain Sampling*

Under the domain sampling model, tests are constructed by selecting a number of measures at random from a large homogeneous pool of items designed to measure a person's "true score" in a particular domain. (The true score is the person's observed score corrected for measurement error.) A number of different tests could be constructed with other random samples of items from this domain's item pool. The correlation between a given test score and the average of all test scores (the reliability index) can be shown (within sampling error) to equal the square root of the correlation of any test score with another test score (the reliability coefficient). The reliability coefficient is an estimate of the ratio of variance in true scores to the variance in observed scores (Nunnally and

Bernstein 1994). The important implication of this property is that items (and tests) are interchangeable if they share roughly the same level of reliability and the selection of items covers the difficulty range of the domain under study. Since items are interchangeable within a clearly defined domain, the psychometric properties of items are more important than specific item content. In fact, items could be selected at random from among those in the item pool shown to have desirable psychometric properties. But something beyond simple random sampling in selection, such as stratifying the pool of reliable items based on difficulty, would help assure the difficulty range requirement is met, for example. One fundamental desirable psychometric property is measurement invariance across groups.

Other methodological developments allow more precise equating of scores across tests with differing subsets of items as long as there is a common core of items (Kolen and Brennan 2004). So, every item does not have to be asked every survey year for scale scores to be equated with some precision if the requirements outlined above are met.

### 5.2 *Conclusion*

Our results help explain why researchers have been frustrated in their attempt to measure and explain apparent knowledge gaps between various grouping variables including participation, media use, educational attainment, income, and age. Models seeking to explain observed differences have been unsuccessful because the construct of political knowledge is apparently qualitatively different between subgroups based on these grouping variables—excluding media use. Attempts to explain these differences will thus be unsuccessful, if political knowledge is measured using the full battery, because the measures are not sufficiently invariant to permit valid comparisons. The inconsistent results may also stem from a lack of conceptual clarity regarding the measurement model underpinning political knowledge. Our results suggest that the use of indexes assumed to be measured without error cannot be supported.

The validity of political knowledge scales should not be established by fiat, and it is apparent that measurement invariance cannot be ruled out in many instances. Thus, we propose the following suggestions to maximize the probability of obtaining measurement invariance. First, research employing scales measuring latent traits should demonstrate invariance between the sub-populations of interest. Otherwise, we cannot tell whether differences found between groups on the scale are products of real differences on the trait of interest or products of measurement artifacts. Second, researchers should follow several guidelines that should facilitate the construction of valid scales. Excessively long batteries may inhibit invariance because they increase the opportunity to include items that are related to group membership after controlling for the latent trait. Thus, researchers should limit their scales to the items that are relevant theoretically to each group under study. Likewise, we echo the previous calls for close consideration of the implications of survey response patterns for scale measurement.

Admittedly, these findings and recommendations pose a dilemma for researchers. It is likely that theoretically important items are not sufficiently invariant across subgroups to be included in a scale. Rather than ignore this threat to validity and forge ahead with them anyway, we would suggest that uncovering indicators that are not invariant across groups be considered an important finding in its own right and be used as a catalyst to search for an explanation as to why this is the case.

The good news is that in the context of knowledge scales, a large number of items may not be needed to capture the range of an individual's knowledge. Delli Carpini and Keeter (1993) find that the five best items can explain 75% of the variance in a thirty-nine-item measure and the ten best items account for 90% of the variance.[18] In our analysis, we find the picture is somewhat more complicated in that items that are invariant for some grouping variables are not invariant for

[18] The five-item subset that Delli Carprini and Keeter (1993) recommend, is from the 1991 ANES Pilot Study, and several of those items are not repeated in the survey years under study here. This analysis is based on a more recently utilized knowledge scale in Dow (2009).

others. Thus, survey instruments should increase the variety of knowledge items even as researchers become more selective in their application. Work can also begin on equating items using IRT models. We concur with the advice of Smiley (1999) and Stolle and Gidengil (2010) to consider including items gauging respondents' ability to obtain and utilize government benefits and services. This additional variety may improve the utility of knowledge scales by providing scholars with a greater selection of items that are theoretically relevant to the sub-populations of interest.

## Funding

National Science Foundation (SBR-9707741, SBR-9317631, SES-9209410, SES-9009379, SES-8808361, SES-8341310, SES-8207580, and SOC77-08885). Any opinions, findings, and conclusions or recommendations expressed in these materials are those of the author(s) and do not necessarily reflect the views of the funding organizations.

## References

Ackerman, Terry A. 1992. A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement* 29(1):67–91.

Ahn, T. K., Robert Huckfeldt, and John Barry Ryan. 2010. Communication, influence, and informational asymmetries among voters. *Political Psychology* 31:763–87.

Alba, Joseph W., and Lynn Hasher. 1983. Is memory schematic? *Psychological Bulletin* 93:203–31.

Alvarez, R. Michael, and Jonathan Nagler. 2004. Party system compactness: Measurement and consequences. *Political Analysis* 12(1):46–62.

American National Election Studies (www.electionstudies.org). 2010. Time Series Cumulative Data File [dataset]. Stanford University and the University of Michigan [producers and distributors].

Arbuckle, James L. 2009. *Amos 18 User's Guide [Computer software]*. Crawfordville, FL: Amos Development Corporation.

Asparouhov, Tihomir, and Bengt Muthen. 2006. *Robust chi square difference testing with mean and variance adjusted test statistics (Web Notes: No. 10, May 26, 2006)*. http://www.statmodel.com/download/webnotes/webnote10.pdf (accessed July 12, 2010).

Axelrod, Robert. 1973. Schema theory: An information processing model of perception and cognition. *American Political Science Review* 67(4):1248–66.

Bagozzi, Richard P. 2011. Measurement and meaning in information systems and organizational research: Methodological and philosophical foundations. *MIS Quarterly* 35(2):261–92.

Bartels, Larry M. 1996. Uninformed votes: Information effects in presidential elections. *American Journal of Political Science* 40(1):194–230.

Blalock, Hubert M. 1964. *Causal inferences in nonexperimental research*. Chapel Hill: University of North Carolina Press. http://www.citeulike.org/group/108/article/106824 (accessed January 16, 2013).

Bollen, Kenneth A. 2011. Evaluating effect, composite and causal indicators in structural equation models. *MIS Quarterly* 35(2):359–72.

Bollen, Kenneth A., and Kwok-Fai Ting. 1993. Confirmatory tetrad analysis. *Sociological Methodology* 23:147–75.

———. 2000. A tetrad test for causal indicators. *Psychological Methods* 5(1):3–22.

Bollen, Kenneth, and Richard Lennox. 1991. Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin* 110(2):305–14.

Bollen, Kenneth A., and Shawn Bauldry. 2011. Three cs in measurement models: Causal indicators, composite indicators and covariates. *Psychological Methods* 16(3):265–84.

Borsboom, Denny. 2006. The attack of the psychometricians. *Psychometrika* 71(3):425–40.

Brown, Timothy. 2006. *Confirmatory factor analysis for applied research*. New York: Gilford Press.

Byrne, Barbara, Richard Shavelson, and Bengt Muthen. 1989. Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin* 105:456–66.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. The statistical analysis of roll call data. *American Political Science Review* 98(02):355–70.

Conover, Pamela Johnston, and Stanley Feldman. 1984. How people organize the political world: A schematic model. *American Journal of Political Science* 28(1):95–126.

Davidov, Eldad. 2009. Measurement equivalence of nationalism and constructive patriotism in the ISSP: 34 countries in a comparative perspective. *Political Analysis* 17(1):64–82.

Davis, Darren, and Brian D. Silver. 2003. Stereotype threat and race of interviewer effects in a survey on political knowledge. *American Journal of Political Science* 47(1):33–45.

Delli Carpini, Michael X., and Scott Keeter. 1993. Measuring political knowledge: Putting first things first. *American Journal of Political Science* 37(4):1179–206.

———. 1996. *What Americans know about politics and why it matters*. New Haven, CT: Yale University Press.

———. 2000. Gender and political knowledge. In *Gender and American politics: Women, men and the political process*, eds. Sue Tolleson-Rinehart and Jyl. J. Josephson, 21–52. Armonk, NY: M.E. Sharpe.

Diamantopoulos, Adamantios, and Heidi M. Winklhofer. 2001. Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research* 38(2):269–77.

Dolan, Kathleen. 2011. Do women and men know different things? Measuring gender differences in political knowledge. *Journal of Politics* 73(01):97–107.

Dow, Jay K. 2009. Gender differences in political knowledge: Distinguishing characteristics-based and returns-based differences. *Political Behavior* 31(1):117–36.

Downs, Anthony. 1957. *An economic theory of democracy*. New York: Harper and Row.

Edwards, Jeffrey R. 2001. Multidimensional constructs in organizational behavior research: An integrative analytical framework. *Organizational Research Methods* 4(14):144–92.

———. 2011. The fallacy of formative measurement. *Organizational Research Methods* 14(2):370–88.

Edwards, Jeffrey R., and Richard P. Bagozzi. 2000. On the nature and direction of relationships between constructs and measures. *Psychological Methods* 5(2):155–74.

Fornell, Claes, and Fred Bookstein. 1982. Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. *Journal of Marketing Research* 19(4):440–52.

Franz, Michael M., Paul B. Freedman, Kenneth M. Goldstein, and Travis N. Ridout. 2007. *Campaign Advertising and American Democracy*. Philadelphia: Temple University Press.

Frazer, Elizabeth, and Kenneth Macdonald. 2003. Sex differences in political knowledge in Britain. *Political Studies* 51(1):67–83.

Freedman, Paul B., Michael M. Franz, and Kenneth M. Goldstein. 2004. Campaign advertising and democratic citizenship. *American Journal of Political Science* 48(4):723–41.

Galston, William A. 2001. Political knowledge, political engagement, and civic education. *Annual Review of Political Science* 4(1):217–34.

Gibson, James L., and Gregory A. Caldeira. 2009. Knowing the Supreme Court? A reconsideration of public ignorance of the High Court. *Journal of Politics* 71(02):429–41.

Hauser, Robert M., and Arthur S. Goldberger. 1971. The treatment of unobservable variables in path analysis. *Sociological Methodology* 3:81–117.

Highton, Benjamin. 2009. Revisiting the relationship between educational attainment and political sophistication. *Journal of Politics* 71(04):1564–76.

Hipp, John R., Daniel J. Bauer, and Kenneth A. Bollen. 2005. Conducting tetrad tests of model fit and contrasts of tetrad-nested models: A new SAS macro. *Structural Equation Modeling: A Multidisciplinary Journal* 12(1): 76–93.

Hipp, John R., and Kenneth A. Bollen. 2003. Model fit in structural equation models with censored, ordinal, and dichotomous variables: Testing vanishing tetrads. *Sociological Methodology* 33(1):267–305.

Horn, John, and John McArdle. 1992. A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research* 18:117–44.

Huckfeldt, Robert. 2001. The social communication of political expertise. *American Journal of Political Science* 45(2):425–38.

Huckfeldt, R., and J. Sprague. 1995. *Citizens, politics, and social communication: Information and influence in an election campaign*. New York: Cambridge University Press.

Iacobucci, Dawn. 2010. Structural equations modeling: Fit indices, sample size, and advanced topics. *Journal of Consumer Psychology* 20(1):90.

Jackman, Simon. 2008. Measurement. In *The Oxford handbook of political methodology*, eds. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier, 119–51. New York: Oxford University Press.

Jöreskog, Karl, and Dag Sörbom. 2006. LISREL 8.8 for Windows [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Kolen, Michael J., and Robert L. Brennan. 2004. *Test equating, scaling and linking: Methods and practices*, 2nd ed. New York: Springer.

Krosnick, Jon A., Arthur Lupia, Matthew DeBell, and Darrell Donakowski. 2008. *Problems with ANES questions measuring political knowledge*. http://www.electionstudies.org/announce/newsltr/20080324PoliticalKnowledgeMemo.pdf (accessed June 22, 2010).

Lambert, Ronald D., James E. Curtis, Barry J. Kay, and Steven D. Brown. 1988. The social sources of political knowledge. *Canadian Journal of Political Science* 21(2):359–74.

Lau, Richard R., and David P. Redlawsk. 2001. Advantages and disadvantages of cognitive heuristics in political decision making. *American Journal of Political Science* 45(4):951–71.

———. 2008. Older but wiser? Effects of age on political cognition. *Journal of Politics* 70(1):168–185.

Lizotte, Mary-Kate, and Andrew H. Sidman. 2009. Explaining the gender gap in political knowledge. *Politics & Gender* 5(2):127–51.

Lodge, Milton, and Charles S. Taber. 2000. Three steps toward a theory of motivated political reasoning. In *Elements of reason: Cognition, choice, and the bounds of rationality*, eds. Arthur Lupia, Matthew D. McCubbins, and Samuel L. Popkin, 183–213. New York, NY: Cambridge University Press.

Lodge, Milton, Kathleen McGraw, and Patrick Stroh. 1989. An impression-driven model of candidate evaluation. *American Political Science Review* 83(2):399–419.

Lupia, Arthur. 1994. Shortcuts versus encyclopedias: Information and voting behavior in California insurance reform elections. *American Political Science Review* 88(1):63–76.

Luskin, Robert C. 1987. Measuring political sophistication. *American Journal of Political Science* 31(4):856–99.

———. 1990. Explaining political sophistication. *Political Behavior* 12(4):331–61.

MacCallum, Robert C., and Michael W. Browne. 1993. The use of causal indicators in covariance structure models: Some practical issues. *Psychological Bulletin* 114(3):533–41.

Macdonald, Stuart Elaine, George Rabinowitz, and Ola Listhaug. 1995. Political sophistication and models of issue voting. *British Journal of Political Science* 25(4):453–83.

MacKenzie, Scott B., Philip M. Podsakoff, and Cheryl Burke Jarvis. 2005. The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *Journal of Applied Psychology* 90(4):710–30.

McGlone, Matthew S., Joshua Aronson, and Diane Kobrynowicz. 2006. Stereotype threat and the gender gap in political knowledge. *Psychology of Women Quarterly* 30(4):392–98.

Meredith, William. 1993. Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58:525–43.

Miller, Warren E., Donald R. Kinder, Steven J. Rosenstone, and the National Election Studies. American national election studies, 1992 time series study [dataset]. Ann Arbor, MI: University of Michigan, Center for Political Studies [producer and distributor], 1999.

Millsap, Roger, and Jenn Yun-Tein. 2004. Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research* 39:479–515.

Mondak, Jeffery J. 1999. Reconsidering the measurement of political knowledge. *Political Analysis* 8(1):57–82.

———. 2001. Developing valid knowledge scales. *American Journal of Political Science* 45(1):224–38.

Mondak, Jeffery J., and Mary R. Anderson. 2004. The knowledge gap: A reexamination of gender-based differences in political knowledge. *Journal of Politics* 66(2):492–512.

Morehouse Mendez, Jeanette, and Tracy Osborn. 2010. Gender and the perception of knowledge in political discussion. *Political Research Quarterly* 63(2):270–80.

Muthen, Bengt, and Anders Christoffersson. 1981. Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika* 46:407–19.

Muthen, Bengt, and Tihomir Asparouhov. 2002. *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus (Mplus Web Notes: No. 4, Version 5, December 9, 2002)*. http://www.statmodel.com/examples/webnote.shtml#web4 (accessed May 6, 2010).

Muthen, Linda, and Bengt Muthen. 2010. *Mplus User's Guide*. 6th ed. (1998–2010). Los Angeles, CA: Muthen & Muthen.

Nunnally Jum, C., H. Ira, and Bernstein. 1994. *Psychometric theory*. 3rd ed. NewYork: McGraw-Hill.

Palfrey, Thomas R., and Keith T. Poole. 1987. The relationship between information, ideology, and voting behavior. *American Journal of Political Science* 31(3):511–30.

Pantoja, Adrian D., and Gary M. Segura. 2003. Fear and loathing in California: Contextual threat and political sophistication among Latino voters. *Political Behavior* 25(3):265–86.

Pietryka, Matthew T., and Randall C. MacIntosh. 2013. *Replication data for: An analysis of ANES items and their use in the construction of political knowledge scales*. http://hdl.handle.net/(1902).1/21218 UNF:5:mZ5Xd6shurv4qyVRk7M4Fg== IQSS Dataverse Network [Distributor] V1 [Version].

Popkin, Samuel L., and Michael A. Dimock. 1999. Political knowledge and citizen competence. In *Citizen competence and democratic institutions*, eds. Stephen L. Elkin and Karol Edward Soltan, 117–46. University Park, PA: Penn State Press.

Prior, Markus. 2005. News vs. entertainment: How increasing media choice widens gaps in political knowledge and turnout. *American Journal of Political Science* 49(3):577–92.

Prior, Markus, and Arthur Lupia. 2008. Money, time, and political knowledge: Distinguishing quick recall and political learning skills. *American Journal of Political Science* 52(1):169–83.

Rapoport, Ronald B. 1979. What they don't know can hurt you. *American Journal of Political Science* 23(4):805–15.

Satorra, Albert. 2000. Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In *Innovations in multivariate statistical analysis: A Festschrift for Heinz Neudecker*, eds. Risto D. H. Heijmans, D. S. G. Pollock, and Albert Satorra, 233–47. New York: Springer. http://www.econ.upf.edu/docs/papers/downloads/395.pdf (accessed July 14, 2010).

Satorra, Albert, and Peter Bentler. 1999. *A scaled difference chi-square test statistic for moment structure analysis*. Technical Report, University of California, Los Angeles. http://www.springerlink.com/content/q4u780616r702384/ (accessed September 15, 2011).

Seeman, Melvin. 1966. Alienation, membership, and political knowledge: A comparative study. *Public Opinion Quarterly* 30(3):353–67.

Shadish, W., T. Cook, and D. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Smiley, M. 1999. Democratic citizenship: A question of competence. In *Citizen competence and democratic institutions*, eds. Stephen L. Elkin and Karol Edward Soltan, 371–83. University Park, PA: Pennsylvania State University Press.

Stegmueller, Daniel. 2011. Apples and oranges? The problem of equivalence in comparative research. *Political Analysis* 19(4): 471–87.

Stolle, Dietlind, and Elisabeth Gidengil. 2010. What do women really know? A gendered analysis of varieties of political knowledge. *Perspectives on Politics* 8(1):93.

Taber, Charles S., and Milton Lodge. 2006. Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science* 50(3):755–69.

The American National Election Studies (www.electionstudies.org). The 1996 time series study [dataset]. Ann Arbor, MI: University of Michigan, Center for Political Studies [producer and distributor].

The American National Election Studies (www.electionstudies.org). The 2000 time series study [dataset]. Ann Arbor, MI: University of Michigan, Center for Political Studies [producer and distributor].

The American National Election Studies (www.electionstudies.org) Time series cumulative data file [dataset]. Stanford University and the University of Michigan [producers and distributors], 2010.

The National Election Studies (www.electionstudies.org). The anes 2004 time series study [dataset]. Ann Arbor, MI: University of Michigan, Center for Political Studies [producer and distributor].

Vandenberg, Robert, and Charles Lance. 2000. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods* 2:4–69.

Verba, Sidney, Kay Lehman Schlozman, and Henry E. Brady. 1995. *Voice and equality: Civic voluntarism in American politics*. Cambridge, MA: Harvard University Press.

Verba, Sidney, Nancy Burns, and Kay Lehman Schlozman. 1997. Knowing and caring about politics: Gender and political engagement. *Journal of Politics* 59(4):1051–72.

Widman, Keith, and Steven Reise. 1997. Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In *The science of prevention: Methodological advances from alcohol and substance abuse research*, eds. K. J. Bryant, M. Windle, and S. G. West, 281–324. Washington, DC: American Psychological Association.

Young, Dannagal Goldthwaite. 2004. Late-night comedy in election 2000: Its influence on candidate trait ratings and the moderating effects of political knowledge and partisanship. *Journal of Broadcasting & Electronic Media* 48(1):1–22.

Yum, June O., and Kathleen E. Kendall. 1988. Sources of political information in a presidential primary campaign. *Journalism Quarterly* 65:148–51.

Zaller, John R. 1992. *The nature and origins of mass opinion*. New York, NY: Cambridge University Press.