Political Communication, 30:620–634, 2013

ISSN: 1058-4609 print / 1091-7675 online
DOI: 10.1080/10584609.2013.819539

Routledge
Taylor & Francis Group

The Workshop

The Challenge of Measuring Media Exposure: Reply to Dilliplane, Goldman, and Mutz

MARKUS PRIOR

Political communication research has long been plagued by inaccurate self-reports of media exposure. Dilliplane, Goldman, and Mutz (2013) propose a new survey-based measure of "televised exposure to politics" that avoids some of the features that lead to self-report error and that has already been adopted by the American National Election Study. Yet the validity of the new measure has not been independently tested. An analysis reveals several weaknesses. First, construct validity of the new measure is low because it does not attempt to measure the amount *of exposure to news programs, news channels, or news overall. Second, its convergent validity is poor by several different criteria. For example, the new measure shows barely any increase in news exposure as the 2008 presidential election approached. Third, the authors' criterion for predictive validity is neither necessary nor sufficient. Dilliplane, Goldman, and Mutz are right that measuring the media exposure of survey respondents in a valid and reliable way is critical for progress in political communication research. But given the inability of many respondents to report their own exposure, it is necessary to monitor the media use of survey respondents automatically.*

Keywords exposure, self-report, validity, news, measurement

People are not very accurate in reporting whether, how often, and for how long they did particular things (e.g., Burton & Blair, 1991; Hadaway, Marler, & Chaves, 1993; Rutherford & Fernie, 2005). Anecdotal evidence has long suggested that self-reports of media exposure, too, suffer from this problem (e.g., Price & Zaller, 1993). Recently, a number of studies have more systematically documented the extent of systematic error in self-reported exposure to television news (Prior, 2009a, 2009b, 2013), political advertising (Ansolabehere & Iyengar, 1998; Vavreck, 2007), and presidential debates (Prior, 2012). These findings call into question that survey-based self-reports can yield valid ordinal (let alone interval or ratio) media exposure scales: It is doubtful that respondents who report high news or campaign exposure

Markus Prior is Associate Professor of Politics and Public Affairs, Woodrow Wilson School and Department of Politics, Princeton University.

I would like to thank Gaurav Sood for research assistance and Doug Arnold, Larry Bartels, Marty Gilens, Andy Guess, Skip Lupia, Thomas Leeper, Tali Mendelberg, Bob Shapiro, Nick Valentino, and Lynn Vavreck for their feedback on earlier versions of this article.

Address correspondence to Markus Prior, Woodrow Wilson School, Princeton University, Princeton, NJ 08544-1013, USA. E-mail: mprior@princeton.edu

actually were exposed to more news or campaign content than respondents who report less exposure.

Dilliplane, Goldman, and Mutz (2013) propose a new survey-based measure of "televised exposure to politics" that has already been adopted by the American National Election Study (ANES) for its 2012 election survey. All survey respondents are first asked "From which of the following sources have you heard anything about the presidential campaign?" Respondents who check "television news programs" or "television talk shows, public affairs or news analysis programs" are then presented with a list of specific programs and asked which of them they "watch regularly on television? Please check any that you watch at least once a month." The authors propose several ways of scoring responses, including an additive scale of the number of programs checked.

In their measurement design, Dilliplane, Goldman, and Mutz avoid some of the features that commonly lead to errors in self-reports. Mindful of the unrealistic demands on respondents involved in recalling and estimating the exact duration of exposure, they only ask respondents to report "regular" viewing, which they define as "at least once a month." Instead of asking about exposure to news in general, exposure to news channels, or exposure to types of programs, their question covers a list of specific programs, so respondents do not have to make their own determination of what constitutes "the news." Taking account of a more diverse media environment, they include not only traditional news formats, but also opinion shows, talk shows, and political comedy.

Yet, the new measure also makes several new assumptions and retains weaknesses of previous self-report questions. Although respondents do not have to determine what constitutes "the news," they have to decide if coverage they saw was "about the presidential campaign." They have to recognize the name of a program in order to report exposure to it. Critically, the new measure assumes that respondents can accurately report whether or not they watch specific programs in a specific time period. While Dilliplane, Goldman, and Mutz are right to note that previous measures that asked respondents to report the frequency or duration of exposure make unrealistic demands, they continue to rely on the assumption that respondents remember past exposure. Given that many people cannot accurately report whether or not they watched a presidential debate (Prior, 2012), probably the most salient example of political television, this assumption seems overly optimistic. Finally, the new measure assumes that the number of different programs a respondent reports watching is a good proxy for "the extent" (Dilliplane et al., 2013, p. 236) of news exposure. This assumption implausibly assigns two respondents who watch the same number of programs the same exposure score even if one of them watches these programs every day while the other watches them once a month.

Developing better measures of media exposure is a pressing goal. For many of the most important questions in the social sciences, it is critical to know what content people are exposed to and for how long. Existing survey questions do not provide valid measures of this key concept. The need for measurement innovation and the quick adoption of the new measure by the ANES make it particularly important to verify that the "program list technique" does in fact "do a superior job of capturing exposure to the varied political content available in today's fragmented media environment" (Dilliplane et al., 2013, p. 236). Dilliplane, Goldman, and Mutz base this claim on two empirical demonstrations: First, they show that the new exposure measure exhibits high levels of reliability. Second, within-person change in exposure predicts within-person change in candidate knowledge, which they consider evidence for predictive validity. Both analyses are conducted using the 2008 National Annenberg Election Survey (NAES) online panel, which included the new questions in Waves 2, 4, and 5.

This article offers a series of challenges to the conclusion that the program list technique improves the measurement of news exposure. Although the authors deserve credit for addressing some of the concerns with self-reports, the new measurement approach exacerbates several other problems. First, the program list technique does not capture what we want to know most about media use because it does not attempt to measure the amount of exposure, only the number of different programs a respondent watches over the course of a month (low construct validity). Second, the correspondence between the new measure and independent measures of the same concept is poor by several different criteria (low convergent validity). Third, the relationship between the new measure and candidate knowledge does not validate the new measure because people can learn about the candidates without watching television news and watch television news without learning about the candidates. Fourth, the reliability of the new measure would only be relevant in conjunction with a compelling demonstration of validity because a highly reliable measure that lacks validity does not improve measurement. This article explains these concerns with the program list technique in greater detail and provides empirical analyses that cast serious doubt on the validity of the new measure.

Regardless of the weaknesses of their new measure, however, Dilliplane, Goldman, and Mutz are right to focus attention on measurement problems in media research. Existing survey-based measures of media exposure lack validity (Ansolabehere & Iyengar, 1998; Price & Zaller, 1993; Prior, 2009a, 2012, 2013; Vavreck, 2007), and efforts to design more valid self-reports (e.g., Althaus & Tewksbury, 2007; Prior, 2009b) may fall short not because researchers are doing poor work, but because respondents cannot report information they do not have and cannot infer. Hence, this note ends with a more radical proposal: If we want to measure the media exposure of our survey respondents, we have to do it without relying on their memory.

Construct Validity

Construct validity describes the correspondence between a measure and the concept (or construct) the measure is intended to capture. The construct that Dilliplane, Goldman, and Mutz aim to measure is "the extent to which people have been exposed to various kinds of political content" on television (p. 236). They break up the measurement task into three parts: a screening question to identify respondents who were not exposed at all, a "program list" to identify regular viewers of specific programs, and the construction of a scale that combines program-specific answers into an overall measure of exposure. Different concerns about construct validity arise at each stage.

The Screen

The screening question identifies respondents who have not "heard anything about the presidential campaign." These respondents are not subsequently asked about exposure to individual programs. This screen creates two problems. First, it redefines the domain of exposure by limiting relevant content to presidential campaign coverage. Even though Dilliplane, Goldman, and Mutz aim to measure "televised exposure to politics," they screen out respondents who encountered political content unrelated to the presidential campaign.

Second, the use of any kind of screening item threatens to introduce systematic measurement error because it requires respondents to determine if the content they encountered was "about the presidential campaign." Respondents who were in fact exposed to campaign coverage may not remember it or fail to report it because they did not appreciate its campaign relevance.

Although both problems could be solved easily by dropping the screen, it is also used in the 2012 ANES.

The Program List

After the initial screen, respondents are presented with a list of programs and asked "Which of the following programs do you watch regularly on television? Please check any that you watch at least once a month." This question thus asks screened respondents to report program exposure regardless of whether they encountered political content or "anything about the presidential campaign" while watching the program. For example, a respondent who watches *Oprah, Ellen*, or *The View* would receive a higher exposure score even if these programs did not cover politics when the respondent tuned in (and even if the respondent is aware of this). Post hoc weighting by the average amount of political content on each program cannot remedy this problem because it assumes that all self-reported viewers of a program were exposed to the same content.

The program list technique imposes a considerable cognitive burden on respondents. They can only answer the question accurately if they recognize the names of the programs they watch. Most of the programs do not identify on which channels they air.[1] Respondents have to remember if they watched *Nightline, Frontline*, or *Dateline*. This requirement might be manageable for well-known and distinctive shows, such as *Oprah* or *The O'Reilly Factor*, but presents a challenge for respondents who typically tune into channels without looking for a particular program. A Fox News viewer might not know if she happened to watch *Fox Report with Shepard Smith, Studio B with Shepard Smith*, or *Your World with Neil Cavuto*. Dilliplane, Goldman, and Mutz provide no evidence for the supposition that "when viewers watch TV, they likely think in terms of the programs they watch rather than in terms of time units . . . or in terms of researcher-defined categories of political programs" (p. 239). If, as seems at least as likely, many news viewers pick by channel, they may not mark programs they watch (because they do not recognize their names) or mark programs they do not watch (because they select programs most easily associated with the channels they watch).

Asking about many different news programs may inadvertently promote overreporting. When respondents estimate the frequency of behaviors, they often rely on a strategy called "decomposition," which involves breaking up a question into more specific parts (Bradburn, Rips, & Shevell, 1987, p. 160; Schwarz & Oyserman, 2001, pp. 138–139). Decomposition tends to increase reported frequencies without increasing their accuracy. For example, Belli, Schwarz, Singer, and Talarico (2000) compared respondents' estimates of the number of long-distance phone calls they made to their phone records. Respondents who were asked in separate questions about phone calls to different cities overreported to a much greater extent than respondents who were asked about all calls in one question.

There is evidence that decomposition also exacerbates overreporting of news exposure. In the 2008 Pew Media Consumption Survey, respondents were randomly assigned to report how often they "watch cable news channels such as CNN, MSNBC, or the Fox News Cable Channel" or asked about their exposure to each cable channel separately. (All questions offered the same four response options: "regularly," "sometimes," "hardly ever," and "never.") In the first condition, 21% said they "never" watched cable news. Asked separately in the second condition, however, only 13% reported "never" watching CNN, MSNBC, and Fox News. By asking about many different programs separately, the program list technique thus risks further inflating exposure estimates.

Scale Construction

The program list technique yields two types of exposure variables. For individual programs, we know if a respondent reported watching a program "at least once a month" or not. As a measure of overall exposure, Dilliplane, Goldman, and Mutz propose to build an additive scale of the number of different programs a respondent reports watching (optionally with weights for program length or amount of political content.) Assuming for a moment that respondents know which programs they watch, know if they watch a program "regularly . . . at least once a month," and report this information accurately, how well can the program list technique distinguish those who watch a lot of news or campaign coverage from those who watch only a little or none? In other words, how well does it measure the *amount* of news exposure or "the *extent* to which people have been exposed to political content" (p. 236, emphasis added)?

For individual programs, the measure does not distinguish different amounts of exposure. A self-reported viewer might watch a program once a month or once a day. This makes it impossible to determine if the impact of program exposure was due to most viewers being affected by exposure or only the most heavily exposed. Neither would it be possible to determine the relative impact of two different programs. For example, if the coefficient for *CBS Evening News* is bigger than the coefficient for *The Daily Show with Jon Stewart* in a model explaining political knowledge, we could not determine whether both shows are equally informative and self-reported viewers of *CBS Evening News* spent more time watching the program than viewers of *The Daily Show with Jon Stewart*, or whether the average minute of *CBS Evening News* is more educational than the average minute of *The Daily Show with Jon Stewart*.

The additive exposure scale (whether weighted by program characteristics or not) does not measure amount of exposure either. The number of programs a person watches at least once a month is a measure of the breadth of her news repertoire. It is not a good ordinal measure of media exposure because it assigns a higher score to someone who watches two different programs once a month (two instances of monthly exposure) than someone who watches the same one program every day (30 instances of monthly exposure).

The concept of news repertoire does not play much of a role in theories of media effects. How many different programs someone watches is not particularly relevant unless we also know the duration of exposure. Instead, the additive scale can easily lead to inaccurate inferences. From their regression of political knowledge on the additive exposure scale, the authors conclude that "increases in exposure to political television significantly predict knowledge gain" (p. 241). What they have actually shown is that an increase in the number of different news programs watched at least once a month is related to knowledge gain. At the same time, the results imply that two otherwise identical respondents who watch the same number of programs will be equally knowledgeable even if one of them watches these programs every day while the other watches them once a month.

Convergent Validity

Convergent validity describes the fit between independent measures of the same underlying concept. According to past research, convergent validity, assessed by the fit between self-reports and independent assessments of a behavior, is low for many non-political behaviors (e.g., Burton & Blair, 1991; Hadaway et al., 1993; Hadaway, Marler, & Chaves, 1998; Presser & Stinson, 1998; Rutherford & Fernie, 2005) as well as for turnout

(e.g., Ansolabehere & Hersh, 2012; Belli, Traugott, & Beckmann, 2001; Bernstein, Chadha, & Montjoy, 2001; Sigelman, 1982; Silver, Anderson, & Abramson, 1986; Traugott & Katosh, 1979) and media exposure (e.g., Price & Zaller, 1993; Prior, 2009a, 2012; Vavreck, 2007).

There is no single cause of self-report error. Mismatch between actual and reported behavior can arise at various stages in the response process (Schwarz, 1999; Schwarz & Oyserman, 2001; Tourangeau, Rips, & Raskinski, 2000). Hence, there are several different reasons to expect that it may be difficult or impossible for respondents to accurately report whether they watch a program "regularly" or "at least once a month." Respondents who do not recognize the name of a program might underreport their exposure. Forgetting is another reason for underreporting, as individual episodes of a behavior are often not retrievable when the behavior is of low salience (e.g., Burton & Blair, 1991). Respondents may incorrectly recall episodes as having occurred during the reference period. This error, called "telescoping" (Schwarz & Oyserman, 2001), would cause overreporting. Flawed estimation can also inflate self-reports (Burton & Blair, 1991; Prior, 2009b).

The weight of past evidence and theory on behavioral self-reports would thus seem to put the burden of proof on proponents of new self-report measures. Yet, Dilliplane, Goldman, and Mutz offer no assessment of convergent validity. In this section, I fill this gap by comparing estimates from the program list technique to independent assessments of news exposure.

Comparing the Program List Technique to Nielsen Data

In the NAES, the percentage of self-reported "Fox News" viewers was 34 in Wave 2 (January 1–March 31, 2008), 35 in Wave 4 (August 29–November 4), and 35 in Wave 5 (November 5, 2008–January 31, 2009).[2] These estimates can be compared to estimates by the Nielsen Company, which has installed "people meters" in a random sample of about 10,000 U.S. households to monitor live and time-shifted television viewing. The Nielsen metric most comparable to self-reports of watching a program or channel "at least once a month" is the "cume," that is, the number of unique viewers who watch a program or channel for at least some time during a defined time period. The use of the word "regularly" makes the NAES definition of exposure ambiguous. How much exposure qualifies as watching a channel or program "regularly . . . at least once a month"?

Few cume estimates are publicly available, and most of them are for cable channels. As Dilliplane, Goldman, and Mutz do not ask about CNN or MSNBC, comparisons have to focus on the Fox News Channel (FNC). According to Nielsen data reported by Stroud (2011, pp. 208–210), about 10% of adults[3] watched FNC for at least an hour over a 2-week period in April 2008, and 13% did so in October 2008. This amount of viewing, which includes commercial breaks, corresponds to just over 4 minutes per day or 2 hours per month. Yet, according to NAES estimates, the FNC audience was roughly three times as large.

Some may prefer "regularly . . . at least once a month" to include viewing durations of less than an hour over 2 weeks. Monthly 60-minute cumes provide a lower qualifying threshold, but are only available for 2010. Assuming the same ratio between the 60-minute cume and the 6-minute cume (which is available for 2008), about 21% to 23% of adults are estimated to have watched at least 60 minutes of FNC per month during the NAES fielding periods—still considerably below NAES estimates.

Naturally, the least overreporting occurs by comparison with the most generous definition of "regular" exposure. Sixty minutes per month—or about 2 minutes on an average

day—is already a low threshold for "regular" viewing. After all, Dilliplane, Goldman, and Mutz aim to measure "regular viewing rather than an isolated chance exposure" (p. 239), assert that "our media measure tap[s] regular exposure" (p. 244), and concede that their measure "inevitably lose[s] information about fleeting, incidental instances of exposure" (p. 245). Nielsen estimates of the 6-minute monthly cume for the Fox News Channel are available for 2008 (Project for Excellence in Journalism, 2009). The percentage of the adult population watching at least 6 minutes per month was 26 during Wave 2 (January–March), 30 during Wave 4 (September/October), and 29 during Wave 5 (November–January 2009). Six minutes of a channel or program per month, including commercials, includes a lot of "fleeting," "isolated chance exposure." And yet, the NAES estimates *still* exceed Nielsen's estimates of the FNC audience.

At first glance, the difference between NAES estimates (34%–35%) and Nielsen's 6-minute monthly cume (26%–30%) might seem small enough to dismiss. But there is a trap in this line of argument: The only way to take this similarity as evidence of convergent validity is to concede the low construct validity of the NAES measure. In order to make automatic tracking estimates and self-reports vaguely resemble each other, we have to count even the most infrequent, incidental viewers—those who watch no more than a few minutes of programming or commercials *per month*—as regular viewers. But now we are left with a measure of "regular" exposure that lumps together any exposure over 5 minutes per month, which is neither the construct Dilliplane, Goldman, and Mutz intend to measure nor a particularly useful variable for media effects analyses.

An analysis of audiences for specific programs included in the NAES list also suggests low convergent validity. Weekly 6-minute cumes for March 22–28, 2008, can provide a rough approximation of how many people watch *Hannity & Colmes* on Fox News, *Fox Report with Shepard Smith, Lou Dobbs Tonight* on CNN, and *The Situation Room* on CNN. The 6-minute cumes for these four programs were 4.6%, 2.9%, 2.8%, and 3.7%, respectively. For example, Nielsen estimates that 4.6% of the population 18 and older watched more than 5 minutes of *Hannity & Colmes* that week (including commercials). Regular viewer estimates by Dilliplane, Goldman, and Mutz for the same four programs in the same week are 9.8%, 8.9%, 7.9%, and 5.3%, respectively. Watching 6 minutes per week of programs that air each weekday is an overly generous approximation of "regular" viewing. Yet, even so, self-reported audiences are on average more than twice as large. Moreover, there is considerable variation in the extent of mismatch: The self-reported audience for *Fox Report with Shepard Smith* is three times higher than its weekly 6-minute cume, while the difference is below 50% for *The Situation Room*.

Different Validation Benchmarks for Convergent Validity

Dilliplane, Goldman, and Mutz dismiss comparisons with Nielsen data because "far from a passive method of assessing television exposure, the people-meter system is another imperfect form of self-report" (p. 244). This statement glosses over the most critical difference between survey-based measures of media exposure and Nielsen estimates: Nielsen's methodology does not rely on participants' memory.

In multi-member Nielsen households, each television set is attached to a meter that automatically records which channel is on, but household members have to indicate the beginning and end of their viewing by pushing a button, so measurement is indeed not entirely passive. In single-member households, however, automatically recorded household-level viewing can be attributed to the single household member with little ambiguity. And analyses of convergent validity in single-member households show

overreporting of news exposure at a similar rate as in multi-member households (Prior, 2009a).

Dilliplane, Goldman, and Mutz are right that Nielsen methodology and data have not been subjected to sufficient scrutiny, so the properties and quality of its sample cannot be independently verified. But Nielsen is not the only company providing observational data on television exposure. Monitoring technology similar to Nielsen's shows considerable overreporting of news exposure in the Netherlands (Wonneberger, Schoenbach, & van Meurs, 2012). Very different monitoring technology also confirms this conclusion: Integrated Media Measurement Incorporated (IMMI, acquired by Arbitron in 2008) automatically tracked media exposure by providing participants with cell phones that pick up radio and TV sound that is matched against a database of programming. This technology measures exposure automatically ("passively") at the individual level and includes out-of-home viewing. Independent researchers have scrutinized IMMI's measurement approach (Jackman, LaCour, Lewis, & Vavreck, 2012). And as IMMI recruits were also asked about their media use in a survey, assessing the validity of self-reports does not require matching of independent samples. Analysis of IMMI data, too, shows considerable overreporting of news exposure in surveys (LaCour, 2012).

Convergent validity can also be assessed without reliance on any external benchmarks by comparing the same exposure item over time. Lack of convergent validity is demonstrated in Prior (2012) by comparing self-reports of exposure to the same event (a presidential debate) on different days in independent random samples (using a rolling cross-section design). If self-reports were globally valid, reported exposure would not depend on the timing of the interview. Yet for most debates between 2000 and 2008, self-reports in the NAES rolling cross-section surveys vary with time, often dropping considerably as time since the debate increases.

One last evaluation of convergent validity cements doubts about the program list technique. The new exposure measure barely picks up one of the clearest features of presidential campaigns: that exposure rises late in the campaign. Among respondents who completed the exposure questions in all three NAES waves, the average number of programs watched is 5.8 in Wave 2, 5.7 in Wave 4, and 5.9 in Wave 5 (using NAES weights). There is no increase between the first quarter of 2008 (Wave 2) and the period between early September and election day (Wave 4) when news audiences typically peak.

Since each NAES wave was conducted as a rolling cross section, time trends within waves can be considered (although the first and last week of each wave have unique sample compositions because respondents who complete a survey promptly differ from those who do not). The highest monthly average occurs in March 2008 with 6.5 programs, and the average of 6.2 in January 2009 still exceeds the average of 5.9 for the month of the election. One has to look at weekly averages to find greater exposure around election day than at other times of the year. The highest weekly value occurs in the week of November 11, with 6.8 programs watched, and is 0.2 higher than the highest Wave 2 week. Panelists interviewed in the week of November 11 (as part of Wave 5) reported on average .81 more programs than during their Wave 4 interview and .68 more than during their Wave 2 interview.

On a scale that has a theoretical range of 0 to 49, a between-person standard deviation of 6, and an interquartile range of 8, .81 is very small. This within-person change in exposure amounts to less than one-seventh of a standard deviation of the exposure measure.

Evidence that the program list technique can detect mounting exposure during the run-up of a momentous election is even more elusive when considering "regular" Fox News viewers. Their share among fully participating panelists was 33 in Wave 2, 32 in Wave 4, and 33 in Wave 5. Monthly NAES estimates are 31% (January), 34% (February), 36%

(March), 31% (September), 33% (October), and 33% (November). By comparison, the average primetime audience for FNC nearly doubled between the first quarter and October of 2008, according to Nielsen.

Convergent Validity: Summary

There is plenty of evidence indicating that self-reports of media exposure generally and the program list technique specifically lack convergent validity. It is not possible to dismiss all of these demonstrations as flawed by pointing to problems with Nielsen data. Other external benchmarks or internal consistency criteria also show lack of convergent validity.

Furthermore, a discussion of how compelling the case for low convergent validity is should not distract from the lack of positive evidence. Numerous behavioral self-reports and all previously evaluated self-reports of news exposure suffer from low convergent validity. Dilliplane, Goldman, and Mutz offer no evidence that the convergent validity of the program list technique is higher. Without such a demonstration, the vague hope that "this time is different" rings hollow.

Predictive Validity

Dilliplane, Goldman, and Mutz argue that "the gold standard for assessing the validity of media exposure is how well a measure predicts political knowledge gains" (p. 238). Based on their finding that the new exposure measure is related to candidate knowledge in both levels and differences, they declare the new measure validated. Yet, their approach raises a number of concerns that challenge this conclusion.

First, if there is a "gold standard" for validation, it is convergent validity, not predictive validity. A measure with high predictive validity but low convergent validity should be and would be dismissed.

Second, by their standard, many of the traditional news exposure measures that Dilliplane, Goldman, and Mutz dismiss do in fact exhibit high predictive validity. A positive association between self-reported TV news exposure and political knowledge has been shown with the use of panel data by Chaffee and Schleuder (1986) and the use of cross-sectional data by, among others, Price (1993), Chang and Krosnick (2002), Prior (2009b), and Tewksbury, Althaus, and Hibbing (2011). If several existing exposure measures meet the validation criterion, that criterion cannot demonstrate the superior predictive validity of the new measure. And if these other exposure measures can be disqualified on other grounds, then predictive validity is not the gold standard.

Third, the predictive validity test incorrectly assumes that television news exposure is necessary and sufficient for political learning. Boudreau and Lupia (2011, p. 173) call validation of political knowledge through a positive association with political interest "problematic" because "a person can be interested in politics without being knowledgeable and can be knowledgeable without being particularly interested." Their concern applies equally to validation of television news exposure: Someone can learn about the candidates without watching television news and watch television news without learning about the candidates.

By using a panel estimator, Dilliplane, Goldman, and Mutz guard against the simplest version of omitted variable bias, the spurious effect of stable predictors of knowledge. They also control for several alternative time-variant causes of political knowledge (Table C3). Some of these controls have little variance, however (e.g., dummy variables measuring

exposure to the presidential campaign in newsmagazines and on the Internet). And the analyses omit several other time-varying predictors of knowledge, including exposure to political advertising, the party conventions, and the presidential debates. If people learn candidate positions from debate exposure, and the number of programs respondents report watching is correlated with debate exposure, then the relationship between the new exposure measure and candidate knowledge is spurious.

The second half of Boudreau and Lupia's caution is equally important. People may watch television news without learning about the candidates. Experimental studies do show that exposure to television news has a positive average treatment effect on political knowledge (e.g., Neuman, Just, & Crigler, 1992). But similar studies also show that some people forget information quickly after exposure (e.g., Lodge, Steenbergen, & Braun, 1995). Dilliplane, Goldman, and Mutz themselves note that "people are exposed to a great deal of content that . . . is not necessarily recalled" and that "a measure of current events knowledge would still conflate what people are exposed to with what people retain from that exposure" (p. 238).

In a natural environment, viewers are more easily distracted and pay less attention, so exposure might produce smaller learning effects than in the lab. Research on long-term learning effects of television exposure and research that aims to overcome the external validity constraints of experimental studies almost always rely on surveys (early classics include Patterson & McClure, 1976; Chaffee & Schleuder, 1986; and Robinson & Levy, 1986, but the list is long). As a result, the true relationship between the extent of television news exposure and political knowledge days or weeks later is subject to considerable uncertainty because evidence for it comes predominantly from just the kind of studies that Dilliplane, Goldman, and Mutz criticize (correctly, in my judgment) for weak measurement.

Research has compellingly linked the amount and type of TV news coverage of a political issue with public knowledge of the issue (e.g., Barabas & Jerit, 2009; Jerit, Barabas, & Bolsen, 2006). But these studies do not examine TV news exposure. Heavily covered issues may lead to better performance on knowledge questions not because more people watch the coverage, but because these issues generate more interpersonal discussion, command greater attention, or are more easily remembered. Since the true individual-level relationship between news exposure and knowledge is thus fairly uncertain, it cannot serve as a reliable benchmark for predictive validation.

Reliability

Statistical reliability describes the extent to which an empirical indicator "consistently measures whatever it measures" (Gay, 1976, p. 92). It can be defined as the square root of the ratio of "true score" variance to the total variance in an observed indicator. The "true score" in this definition is the expected value of a (hypothetical) set of infinite replications of the measurement for a particular individual. "True score" is thus a statistical concept, not an epistemological declaration. Reliability is independent of validity, and high reliability cannot substitute for a demonstration of validity. Even a measure with low validity can exhibit high true-score reliability.

Dilliplane, Goldman, and Mutz's reliability analysis rests on a number of restrictive assumptions that are standard for panel data with three waves (see the appendix). For the purpose of assessing the reliability of self-reported news exposure, the assumption of uncorrelated measurement error is particularly critical. This assumption likely causes upward bias in reliability estimates. As Zaller (2002, pp. 312–313) explains:

> Over-report bias, which no doubt varies across respondents and items, does not reduce calculated estimates of scale reliability. On the contrary, it tends to enhance apparent reliability. As long as survey respondents exaggerate with some degree of consistency from one exposure item to the next and one survey to the next, over-report bias cuts randomness and thereby enhances estimates of reliability.

There are several reasons why respondents may make similar mistakes in different waves. Measurement error may be correlated across waves if respondents are uncertain about, or confused by, the same elements of the question in repeated waves. Respondents who cannot perfectly recall their exposure may use similar flawed estimation rules to come up with an answer each time they are asked (Prior, 2009b), thus generating serial error correlation. Respondents who feel social desirability pressures may exaggerate their exposure in each wave. If errors are indeed correlated over time, neither the Heise (1969) model employed by Dilliplane, Goldman, and Mutz nor the more commonly used Wiley and Wiley (1970) model will provide appropriate reliability estimates.

Conclusion

Attention to measurement problems in media research is badly needed, so Dilliplane, Goldman, and Mutz's initiative to develop an improved measure is welcome. They rightly aim to simplify the task respondents face when answering exposure questions. Many people cannot recall all instances of media exposure, so they have to estimate exposure instead, which they often do poorly (see Prior, 2009b). Unfortunately, by shifting the measured construct from the amount of news exposure to the number of different programs respondents watch, they end up measuring a concept that is of little theoretical relevance. They are right to criticize general exposure measures that fail to explain what constitutes "the news." However, their alternative is unrealistic in requiring respondents to recognize the names of programs they watch and counterproductive in risking even greater overreporting through decomposition.

Dilliplane, Goldman, and Mutz emphasize the high reliability estimates they obtain for their exposure measures but do not account for the danger Zaller (2002, p. 313) warned about: "Over-report bias can . . . hide the damage it does behind exaggerated reliability estimates." Moreover, in light of doubtful construct and convergent validity, high reliability only tells us that we are reliably measuring something other than the thing we want to measure.

The authors' demonstration of predictive validity assumes that television news exposure is related to political knowledge over periods of days and weeks. This assumption is questionable because it rests mostly on research that draws on self-reports of news exposure. Accepting the assumption that underlies their predictive validity analysis requires accepting that the programs respondents watched actually covered candidate positions just when they were watching, that they paid enough attention to learn the candidates' positions, that they did not quickly forget what they learned, and that they did not learn it from exposure to debates, campaign ads, or other people. How can we know all this so well when our exposure measures are so poor?

A variety of convergent validity tests indicate a clear mismatch between the estimates based on the program list technique and independent assessments of news exposure.

Considerably more people report watching Fox News "at least once a month" than tune in according to Nielsen. For example, the share of people who watch FNC for more

than 60 minutes in a 2-week period is about a third of the share estimated by Dilliplane, Goldman, and Mutz to be "regular" viewers. Similar results emerge for specific programs. The program list technique thus widely overstates the share of the population that watches a program for more than a few minutes. It does not provide much help for future research because it lumps together non-viewers, many individuals who watch a few minutes of news per month, and some who watch several hours per day. Other monitoring technologies and internal inconsistencies corroborate the low convergent validity.

Perhaps most implausibly, Dilliplane, Goldman, and Mutz's measure shows barely any increase in news exposure as the 2008 presidential election approached. The authors acknowledge this, summarizing that "people tended to watch the same number of political programs across the three waves; for the most part, they even stuck with the exact same programs" (p. 246). Whether or not that statement is accurate, the new exposure measure fails to measure the theoretically most consequential behavior: How often and for how long did people watch those programs? Presumably more often and for longer as the election approached, but the program list technique cannot tell us.

It should be obvious that this critique is not a defense of other self-reported news exposure measures. Asking respondents to report the number of days per week they watch network news or listen to NPR shows considerable overreporting (Price, 1993; Prior, 2009a), as do questions about "regular" exposure to cable networks (Prior, 2013), watching presidential debates (Prior, 2012), and exposure to campaign advertising (Ansolabehere & Iyengar, 1998; Vavreck, 2007). The program list technique is a new variation of self-report, so some might hope that "this time is different." The evidence presented suggests, however, that the new measure has flaws that are inherent to any measurement of media exposure that relies on self-reports. In light of these conceptual and empirical problems, the program list technique—or any other memory-based approach—is not recommended for measuring news exposure. Just like users of older ANES data sets should be mindful of the shortcomings of traditional exposure measures, users of the 2012 American National Election Study should think twice before using the new exposure measure.

Dilliplane, Goldman, and Mutz are right that measuring the media exposure of survey respondents in a valid and reliable way is a critical requirement for progress on many important research questions. Yet our continued inability to rely on respondents' own reports leads to a fairly radical proposal: We must automatically monitor the media use of our survey respondents.

Technologies developed by audience measurement companies provide models for such an innovation. For specific platforms, it is already possible to monitor survey respondents' exposure passively (e.g., Gentzkow & Shapiro, 2011; Jackman et al., 2012; LaCour, 2012). Audience research firms, including Nielsen, Arbitron, and comScore, measure television, radio, and Internet use without reliance on self-reports. Cable and satellite providers track media use passively and automatically. Tracking technologies will become cheaper. Rising penetration of digital media and continuing digital convergence will reduce the need to monitor offline media. For social science to take advantage of these technological advances, the most important challenge will be to gain respondents' trust and cooperation, so they agree to share their data.

If we are lucky, the combination of respondents' media use records and their answers to our exposure questions might allow us to calibrate a sufficiently accurate measurement model so that we can correct self-reports when independent monitoring is not available. Since over- and misreporting apparently depend on individual characteristics in complex ways (Prior, 2009a, 2012; Zaller, 2002), the prospect of a validated measurement model may be too optimistic, thus requiring audience tracking for every new survey (focused on

media effects). Even the lucky scenario is very expensive and requires stronger respondent cooperation than current practices. But the alternative is to continue without any compelling measure at all of one of the most central concepts in all social science.

Notes

1. At least one format on the list, "Fox News," is in fact a channel. "CNN Newsroom/Headline News" could be interpreted as a channel, and "MSNBC live" might be understood as watching any live coverage of events on MSNBC or watching the channel live rather than tape-delayed.

2. These estimates include all available respondents regardless of how many other waves they completed and use the weights provided by the NAES.

3. Nielsen estimates of FNC audiences used in this article are for the entire population (P2+). When expressing them as a percentage of the adult population (P18+), I thus assume that FNC viewing in the 2–17 age group is minimal. To the extent that the assumption is incorrect, the adult FNC audience is even smaller.

References

Althaus, S. L., & Tewksbury, D. (2007). *Toward a new generation of media use measures for the ANES*. Retrieved from http://www.electionstudies.org/resources/papers/Pilot2006/nes011903.pdf

Ansolabehere, S., & Hersh, E. (2012). Validation: What big data reveal about survey misreporting and the real electorate. *Political Analysis*, *20*, 437–459.

Ansolabehere, S., & Iyengar, S. (1998). *Message forgotten: Misreporting in surveys and the bias toward minimal effects*. Unpublished manuscript.

Barabas, J., & Jerit, J. (2009). Estimating the causal effects of media coverage on policy-specific knowledge. *American Journal of Political Science*, *53*, 73–89.

Belli, R. F., Schwarz, N. Singer, E., & Talarico, J. (2000). Decomposition can harm the accuracy of behavioral frequency reports. *Applied Cognitive Psychology*, *14*, 295–308.

Belli, R. F., Traugott, M., & Beckmann, M. (2001). What leads to voting overreports? Contrasts of overreporters to validated voters and admitted nonvoters in the American National Election Studies. *Journal of Official Statistics*, *17*, 479–498.

Bernstein, R., Chadha, A., & Montjoy, R. (2001). Overreporting voting: Why it happens and why it matters. *Public Opinion Quarterly*, *65*, 22–44.

Boudreau, C., & Lupia, A. (2011). Political knowlege. In J. N. Druckman, D. P. Green, J. H. Kuklinski, & A. Lupia (Eds.), *Cambridge handbook of experimental political science* (pp. 171–183). New York, NY: Cambridge University Press.

Bradburn, N. M., Rips, L, & Shevell, S. (1987, April 10,). Answering autobiographical questions: The impact of memory and inference on surveys. *Science*, pp. 157–161.

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

Chaffee, S. H., & Schleuder, J. (1986). Measurement and effects of attention to media news. *Human Communication Research*, *13*, 76–107.

Chang, L., & Krosnick, J. (2002). Measuring the frequency of regular behaviors: Comparing the typical week to the past week. *Sociological Methodology*, *33*, 55–80.

Dilliplane, S., Goldman, S., & Mutz, D. (2013). Televised exposure to politics: New measures for a fragmented media environment. *American Journal of Political Science*, *57*, 236–248.

Gay, L. R. (1976). *Educational research: Competencies for analysis and application*. New York, NY: Merrill.

Gentzkow, M., & Shapiro, J. (2011). Ideological segregation online and offline. *Quarterly Journal of Economics*, *126*, 1799–1839.

Hadaway, C., Marler, P., & Chaves, M. (1993). What the polls don't show: A closer look at U.S. church attendance. *American Sociological Review*, *58*, 741–752.

Hadaway, C., Marler, P., & Chaves, M. (1998). Overreporting church attendance in America: Evidence that demands the same verdict. *American Sociological Review*, *63*, 122–130.

Heise, D. R. (1969). Separating reliability and stability in test-retest-correlations. *American Sociological Review*, *34*, 93–101.

Jackman, S. D., LaCour, M., Lewis, J., & Vavreck, L. (2012). *Digital fingerprints: A new method for measuring political advertising*. Unpublished manuscript.

Jerit, J., Barabas, J., & Bolsen, T. (2006). Citizens, knowledge, and the information environment. *American Journal of Political Science*, *50*, 266–282.

LaCour, M. J. (2012). *A balanced news diet, not selective exposure: Evidence from a direct measure of media exposure*. Unpublished manuscript.

Lodge, M., Steenbergen, M., & Braun, S. (1995). The responsive voter: Campaign information and the dynamics of candidate evaluation. *American Political Science Review*, *89*, 309–326.

Neuman, W., Just, M., & Crigler, A. (1992). *Common knowledge: News and the construction of political meaning*. Chicago, IL: University of Chicago Press.

Patterson, T. E., & McClure, R. (1976). *The unseeing eye: The myth of television power in national politics*. New York, NY: Putnam.

Presser, S., & Stinson, L. (1998). Data collection mode and social desirability bias in self-reported religious attendance. *American Sociological Review*, *63*, 137–145.

Price, V. (1993). The impact of varying reference periods in survey questions about media use. *Journalism Quarterly*, *70*, 615–627.

Price, V., & Zaller, J. (1993). Who gets the news? Alternative measures of news reception and their implications for research. *Public Opinion Quarterly*, *57*, 133–164.

Prior, M. (2009a). The immensely inflated news audience: Assessing bias in self-reported news exposure. *Public Opinion Quarterly*, *73*, 130–143.

Prior, M. (2009b). Improving media effects research through better measurement of news exposure. *Journal of Politics*, *71*, 893–908.

Prior, M. (2012). Who watches presidential debates? Measurement problems in campaign effects research. *Public Opinion Quarterly*, *76*, 350–363.

Prior, M. (2013). Media and political polarization. *Annual Review of Political Science*, *16*, 101–127.

Project for Excellence in Journalism. (2009). *The state of the news media 2009*. Washington, DC: Author.

Robinson, J. P., & Levy, M. (1986). *The main source*. Beverly Hills, CA: Sage.

Rutherford, A., & Fernie, G. (2005). The accuracy of footballers' frequency estimates of their own football heading. *Applied Cognitive Psychology*, *19*, 477–487.

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

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication and questionnaire construction. *American Journal of Evaluation*, *22*, 127–160.

Sigelman, L. (1982). The nonvoting voter in voting research. *American Journal of Political Science*, *26*, 47–56.

Silver, B. D., Anderson, B., & Abramson, P. (1986). Who overreports voting? *American Political Science Review*, *80*, 613–624.

Stroud, N. J. (2011). *Niche news: The politics of news choice*. New York, NY: Oxford University Press.

Tewksbury, D., Althaus, S., & Hibbing, M. (2011). Estimating self-reported news exposure across and within typical days: Should surveys use more refined measures? *Communication Methods and Measures*, *5*, 311–328.

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

Traugott, M. W., & Katosh, J. (1979). Response validity in surveys of voting behavior. *Public Opinion Quarterly*, *43*, 359–377.

Vavreck, L. (2007). The exaggerated effects of advertising on turnout: The dangers of self-reports. *Quarterly Journal of Political Science*, *2*, 287–305.

Werts, C. E., Jöreskog, K., & Linn, R. (1971). Comment on the estimation of measurement error in panel data. *American Sociological Review*, *36*, 110–113.

Wiley, D. E., & Wiley, J. (1970). The estimation of measurement error in panel data. *American Sociological Review*, *35*, 112–117.

Wonneberger, A., Schoenbach, K., & van Meurs, L. (2012). Dimensionality of TV news exposure: Mapping news viewing behavior with people-meter data. *International Journal of Public Opinion Research*, *25*, 87–107.

Zaller, J. (2002). The statistical power of election studies to detect media exposure effects in political campaigns. *Electoral Studies*, *21*, 297–329.

Appendix: Reliability Analysis Using Panel Data

In the classical measurement error model, the observed response at wave *t* is a function of the true score π_t and measurement error ε:

$$y_t = \pi_t + \varepsilon_t. \tag{1}$$

The error terms are assumed to have mean zero and variance $\sigma^2_{\varepsilon_t}$ and to be uncorrelated with the latent variables [$E(\varepsilon_t, \pi_t) = 0$].

The relationship between true scores at different times is expressed in the structural part of the model. In a Markov or lag-1 model, the true score at wave *t* depends only on the true score in the preceding wave and a structural disturbance term δ_t. For a three-wave panel, the structural part is

$$\begin{aligned} \pi_1 &= \delta_1(2) \\ \pi_2 &= \beta_{21} \cdots \pi_1 + \delta_2 \\ \pi_3 &= \beta_{32} \cdots \pi_2 + \delta_3. \end{aligned} \tag{2}$$

The β coefficients represent the stability of the true scores between subsequent waves. The disturbance terms δ_t are assumed to have mean zero and variance $\sigma^2_{\delta_t}$ and to be uncorrelated with each other [$E(\delta_t, \delta_s) = 0$, for $t \neq s$], with the true scores in previous waves [$E(\delta_t, \pi_s) = 0$, for $t > s$], and with the error terms [$E(\varepsilon_t, \delta_t) = 0$].

With three panel waves, additional assumptions are necessary to identify the model. Error terms are typically assumed to be serially uncorrelated [$E(\varepsilon_t, \varepsilon_s) = 0$, for $t \neq s$]. Heise (1969) and Wiley and Wiley (1970) use slightly different ways to achieve the necessary remaining identification assumption. Heise (1969) proposes a standardized model (of correlations) with reliabilities assumed to be equal across panel waves. Wiley and Wiley (1970) show that Heise's assumption of equal reliabilities is implausible in most cases and that standardization does not take advantage of all available information (see also Werts, Jöreskog, & Linn, 1971). They introduce an unstandardized model (of variances and covariances) that allows reliabilities to vary but assumes measurement error variances to be constant over time ($\sigma^2_{\varepsilon 1} = \sigma^2_{\varepsilon 2} = \sigma^2_{\varepsilon 3} = \sigma^2_{\varepsilon}$). Contrary to Dilliplane, Goldman, and Mutz's assertion, the Wiley and Wiley (1970) method thus does not "rel[y] on additional assumptions" and is not "less robust than the Heise approach" (p. 240).