Faithful Measures: 
Developing Improved Measures of Religion

Roger Finke, Christopher D. Bader and Edward C. Polson
Roger Finke, Penn State University
Christopher D. Bader, Baylor University
Edward C. Polson, Louisiana State University-Shreveport

www.theARDA.com
ABSTRACT

Despite an avalanche of innovations in conducting data analysis, the measures used to generate the data have undergone remarkably modest change. This is a significant problem for the social scientific study of religion, given the changing substantive and theoretical issues being addressed and the increasing diversity of religions being studied. This paper addresses three areas. First, we identify three sources of measurement error and identify systematic errors as the most severe form, resulting in both Type I and Type II errors when testing hypotheses. Second, we target specific areas where improved measures are needed, addressing both the mundane issues of measurement and the more theoretical questions of measurement design. Third, we propose an agenda and strategy for developing and testing new measures.
Over the past four decades the social sciences have made remarkable strides in data analysis. As late as the 1960s, bivariate tables with a single summary statistic were standard fare, even for major journals. By the 1970s, however, multivariate models had become the new standard and a wide range of new statistics for categorical variables, latent variables, panel studies, path models and others all began to emerge (see Goodman 1972; Duncan 1975; Alwin and Hauser 1975). During the 1980s, many of these new techniques became available to the broader research community for the first time. Fueled by the ever-increasing power and reduced cost of computers, statistical iterations once considered unthinkable were now computed in seconds.

Despite the avalanche of innovations in data analysis, the measures used to generate data have undergone remarkably modest change.¹ Multiple new statistical techniques have been introduced to address correlated error terms, impute values for missing data or to address other measurement violations of statistical assumptions. The measures, however, have received less attention. In part, this is due to a desire to offer comparable measures over time and across surveys, leading principal investigators to replicate previous items.² No doubt the costs of conducting surveys also reduce opportunities for introducing new measures. At a time when the costs of calculating statistics and accessing previous data collections have plummeted, the costs of conducting new surveys continue to rise, and the amount of work involved has shown little change.

But the desire to maintain continuity and avoid costs is only part of the story. Even when new measures are introduced they frequently receive limited evaluation. This lag in measurement developments raises an obvious concern. Even the most sophisticated statistical models cannot change the quality of the data used. In the pithy vernacular of computer scientists: “garbage in, garbage out.”³

Perhaps no other substantive area faces greater measurement challenges than religion.³ Whether defining theoretical concepts, identifying indicators of those concepts or choosing specific questions and responses to serve as measures of the concepts, each step of the measurement process is filled with hurdles that threaten to introduce error. This paper begins with a general discussion of measurement errors and the implications these errors hold for research and theory on religion. Next, we target specific areas where improved measures are needed; addressing both the mundane issues and the more theoretical questions of measurement design. Finally, we offer a proposal for developing and testing new measures. Along the way we will point out potential measurement errors in many data collections. This should not suggest that the studies were poorly conducted or that we view them as studies to be ignored. On the contrary, we are selecting data collections and research studies that warrant attention and should be built upon.

**Measurement Errors: Implications for Theory and Research**

Standard texts on research methods identify two key measurement concerns: reliability and validity. Reliability is the extent to which measures consistently give the same value over multiple observations, pointing to the importance of consistency. Validity is the extent to which an item measures the concept meant to be measured, highlighting the accuracy of the measure (Babbie 2001). Thus, each is concerned with the extent of error introduced through the measures.

Focusing only on reliability and validity, however, often distracts from a more significant concern. How do faulty measures distort or bias research results? More specifically, when do measurement errors distort our overview of a substantive area and when do they lead to a faulty acceptance or rejection of a hypothesis? Some errors distort descriptive profiles, but result in few changes to the relationships being tested. Others bias the relationships being tested, but result in less distortion of the descriptive results. Each of these errors raises different concerns. Specific concerns will vary based on the type of research being conducted.

¹We recognize, of course, the important work of Duane Alwin (2007), Stanley Presser et al. (2004) and others addressing measurement issues in survey design. We argue, however, that measurement issues have received far less attention than data analysis.
²For widely utilized national surveys such as the General Social Surveys (GSS) and the National Election Studies (NES) changes in question wording or response categories compromise the ability to chart trends over time.
³In the early 1990s, a group of scholars exploring the connections between religion and politics advocated for the use of a set of improved measures of religious belief, behavior and belonging. A few were integrated into the NES. For a discussion, see Wald and Smidt (1993).
Moving beyond reliability and validity, we focus on three common types of measurement error: constant, random, and systematic. For each type of error, we discuss the implications for research on religion, offer examples from existing research and model the effects of the error.

**Constant Errors: Less Than Meets the Eye**

Constant measurement errors appear to be the most egregious because they are so obvious. Like an odometer that records 150 miles for every 100 miles driven, responses to some survey items are consistently inflated or deflated. One potential source of constant error in surveys is the overreporting of socially desirable behavior. Public opinion researchers have long known that respondents overreport normative behavior and underreport activities perceived as being deviant or risky (Bradburn 1983; Presser and Traugott 1992; Silver et al. 1986). For instance, researchers examining the political participation of Americans agree that voting in the U.S. is consistently inflated, sometimes by as much as 20 to 30 percent (Abelson et al. 1992; Bernstein et al. 2001; Katosh and Traugott 1981; Presser 1990; Sigelman 1982). And research has shown that the desire to give the socially desirable or "right" answer goes well beyond voting (Axinn and Pearce 2006).

This inflation of socially desirable behavior is clearly evident in religion measures. Perhaps the most extensively documented example is church attendance. In 1993 Kirk Hadaway, Penny Marler and Mark Chaves published an article in the *American Sociological Review (ASR)* that stated U.S. survey respondents overreport church attendance. A series of articles, including a symposium in ASR, soon debated the extent of the overreporting and what the true attendance rate should be. Some claimed that while overreporting is a problem, the percentage of respondents who overreport is not as large as Hadaway et al. contend (Hout and Greeley 1998). Robert Woodberry (1998) suggested that inflated church attendance may be the result of sampling error rather than measurement error. Because surveys tend to oversample churchgoers, they are likely to reveal a higher percentage of weekly church attendees than may be present in the population. In the end, what did we learn from this research? How should it change the way we use the measure of church attendance?

All of the research agrees that church attendance, like other forms of socially desirable behavior, is overreported. Whether the overreporting is twice the actual rate as reported by Hadaway et al. (1993), or only 1.1 times the actual rate as reported by Hout and Greeley (1998), we know that the percentage of the population attending church each week is lower than the percentage found by most surveys. Thus, when profiling rates of church attendance, researchers need to acknowledge that attendance is lower than reported. But our greater concern is the implication that this error holds for research. Does this type of error distort or bias findings? How does it influence relationships between variables. Does it cause us to inappropriately accept or reject a hypothesis?

To the extent that these errors in reporting are constant across the sample (the concern of past research) the summary coefficients of a regression model will, in fact, show no change. But, in our opinion, the concern over constant error in church attendance figures is itself inflated. Constant error introduces little or no bias into the relationship between variables and the strength of relationships remain the same. Church attendance figures are rarely used to forecast the exact number of people attending worship services on a given weekend. Attendance measures are more typically used for hypothesis testing or trend analysis. Such analyses are not distorted by constant errors (see Mckinnon 2003, Orenstein 2002, Smith 2003 and Wald et al. 1993).

To demonstrate this effect, we model a constant error using data from the Baylor Religion Survey 2005. We simulate a constant error by adding a value of one to each respondent’s answer, thereby further

---

4Studies reveal that deviant behaviors, such as substance abuse and theft, are consistently underreported on surveys (Mensch and Kandel 1988; see Wentland and Smith 1993). Also, some research has shown that the level of underreporting for juveniles varies by sex, race and social class (Hindelang, Hirschi and Weiss 1981).

5Along with the six essays included in the ASR symposium, Hadaway, Marler, and Chaves each published additional articles on this topic (Chaves and Cavendish 1994; Hadaway and Marler 1998, 2005; Marler and Hadaway 1999).

6The Baylor Religion Survey 2005 was administered to a random, national sample of U.S. citizens by the Gallup Organization in the fall of 2005. For full details of the sample and methods behind the survey see Bader, Mencken and Froese (2007).
inflating each respondent’s reported church attendance. Figure 1 offers a visual representation of the relationships between biblical literalism and each of our measures of church attendance. Line 1 represents the church attendance and biblical literalism relationship before the constant error was added. Line 2 is the relationship when a value of 1 is added to each respondent’s answer. As reviewed in most entry level statistics books, adding a constant value to all cases has no affect on the correlations and in a regression equation the only difference between the two models is that the Y-intercepts differ by exactly 1 — the value used for our constant error.

Figure 1: Change in Hypothetical Regression Lines with Addition of a Constant

Thus, despite all of the obvious problems constant errors create for descriptive purposes, the errors pose no threat for accurate hypothesis testing. In the case of church attendance, a vast amount of research has been devoted to uncovering the exact amount of error in reporting, but if the error is constant it is of no concern for theory testing. The most significant concern is not if church attendance is inflated, but rather whether that inflation is constant across groups and over time. Does the level of inflation vary by race, social class, region, year or other variables of interest? This is an area that researchers need to explore further.

If constant errors are often less problematic than they appear, random and especially systematic measurement errors are just the opposite. They are often subtle, yet they pose a serious threat for accurate hypothesis testing.

Random Errors: Attenuating Relationships

Whereas constant errors refer to measurement errors that are distributed equally across all cases, random errors refer to errors being distributed randomly or unequally across cases. The misreporting can vary in both amount and direction. Rather than everyone over reporting some behavior, like voting or church attendance, respondents might provide inaccurate or imprecise values for a particular measure, and these errors are then randomly distributed across the cases. For instance, when asked to report their annual income some individuals

7The church attendance item in the BRS has nine possible responses from 1 (Never) to 9 (Several times a week). Adding a constant error of “1” creates a new category of church attendance "10" which does not exist in the original question. For the abstract purpose of this example the issue is of little concern.
will inflate their level of income while others will deflate it (Bound et al. 1990; Kormendi 1988; Withey 1954). Respondents may have difficulty recalling actual income, or they may intentionally misreport earnings. Regardless, the result is the introduction of random error into a measure, a common problem in the social sciences.

How does this type of error impact research results? When compared to constant error, the implications of random error are nearly reversed. Measures impacted by random error offer a descriptive profile that remains true, while the relationships between variables are distorted. They are attenuation though not biased. For example, if the extent of underreporting and overreporting of church attendance is randomly distributed in amount and direction, the statistical mean for the total sample will remain unchanged. Therefore, if a survey is being used to describe the level of religious participation, random error will not distort the results.

The most serious concern is that statistical relationships will be attenuated with the introduction of random errors. In other words, as random errors increase, relationships will be reduced. Because the measure is less accurate for each case and the overall standard errors will increase, random errors will reduce the size and significance of statistical relationships. The end result is that random errors may lead to Type II errors: finding no significant relationship where one actually exists. This type of error is difficult to identify because the true values of measures are often not observable.

Returning to our hypothetical example of church attendance, we can model the effect of random error. For instance, the relationship between age and church attendance is well documented, with older people typically attending church at a higher rate than younger people (Firebaugh and Harley 1991; Hout and Greeley 1987). Using the BRS data we run an OLS model regressing church attendance on a standard set of demographic variables.8 The findings hold few surprises (see Model 1 in Table 1). Males report less frequent attendance than females. Those that are married report higher levels of attendance than those who are not. Evangelicals and Black Protestants report the highest levels of church attendance. And, as other research has found, older individuals report attending church more frequently (b = .067, p < .01).

Next, we introduce random error by using a random-number generator to change the reported ages of BRS respondents. In doing so, we assumed that some respondents would accurately report their ages, while others might underreport or overreport their age by up to 10 years. We also assumed that this process would occur entirely at random. In other words, for every respondent who might overreport his or her age there is another respondent who will underreport his or her age by the same amount. Therefore, we created a modified age variable that randomly increased or decreased a respondent’s reported age by up to 10 years.9 Running our OLS model again, but replacing the original age with our modified age variable we received a coefficient for age that was slightly reduced, but remained significant (see Model 2). When we double the random error, however, allowing up to 20 years of error, the coefficient for age now dropped to .015 and was insignificant (see Model 3). For both models the strength, direction and significance of the other coefficients in the model remain largely unchanged. Although not shown in Table 1, the statistical mean for age showed little change as random errors were introduced. When random errors of up to 10 years were introduced the mean age was near 50 and showed little change when the random error was allowed to increase up to 20 years (49.9). Both are nearly identical to the mean for the unmodified age variable (49.8) and the median of 49 was the same for all three measures.

Table 1: OLS Regression on Church Attendance on Demographic Characteristics and Religious Tradition
(Standardized Coefficients Presented).

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>-.102**</td>
<td>-.101**</td>
<td>-.100**</td>
</tr>
<tr>
<td>Married</td>
<td>.154**</td>
<td>.154**</td>
<td>.159**</td>
</tr>
</tbody>
</table>

8 The variables were coded as follows: gender (male = 1), marital status (married = 1), race (white = 1), family income, age, and religious tradition (Catholic, Black Protestant, Mainline, Jewish, ”other” and no religion with Evangelical as the contrast category).

9 Using Visual Basic we generated a random number for each case between 1 and 21. Respondents who received a number between 1 and 10 had that number subtracted from their age (age – x). Those who received an 11 had no change to their age. Those who received a score between 12 and 21 had x-11 added to their age. For example, a respondent of age 50 with the random number 2, has a modified age of 48. A respondent of age 50 with the random number 13 has a modified age of 52 (50 + (13-11) = 52).
Although some random errors are difficult to identify or avoid, others are unintentionally introduced by researchers. For example, the common practice of collapsing continuous variables, such as age or income, is a frequent source of random error. Drawing again on BRS data Table 2 illustrates what happens to the statistical relationship between age and several measures of religious practice when respondents’ ages are recoded into a few ordinal categories.\(^{10}\) Though still significant, the correlations that age holds with these important religion variables are reduced. Like the random errors reviewed in our simulation this attenuates the statistical relationships being tested and contributes to Type II errors.

These two examples reinforce several important points about random errors. First, because the means show little change when random error is introduced, even variables high in random error can produce accurate descriptive measures. Second, random errors attenuate relationships and can lead to Type II errors. For example, based on the results of Model 3 in Table 1, a researcher would conclude that age does not impact church attendance. Third, because the error is random, and is not related to other variables in the model, it does not significantly alter the other coefficients in the equation. For theory testing this means that the random errors will not bias other relationships in the theoretical model. Fourth, it takes very large random errors to cause Type II errors. Nevertheless, when compared to constant errors, random errors do reduce the accuracy of hypothesis testing.

Table 2: Bivariate correlations between age and religious practice using continuous and ordinal age variables

<table>
<thead>
<tr>
<th>Religious Service Attendance (N = 1673)</th>
<th>Continuous Age Variable</th>
<th>Ordinal Age Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency of Prayer / Meditation (N = 1671)</td>
<td>.124</td>
<td>.112</td>
</tr>
<tr>
<td>Frequency of Reading Sacred Texts (N = 1672)</td>
<td>.130</td>
<td>.125</td>
</tr>
</tbody>
</table>

The examples just given assume that we know that the measurement errors are random, an assumption often made by multivariate statistical models as well. But seldom do we know if error is truly random. In the case of age, the errors could be related to race, gender, education, or age itself. What we do know, however, is

\(^{10}\)The BRS continuous age variable was recoded into the following categories: 18-24, 25-30, 31-44, 45-64, and 65 and over.
that when measurement errors are systematically related to other variables in the model, this introduces multiple problems for the researcher.

Systematic Errors: Increasing Bias

For hypothesis testing, the most severe measurement errors are those that are systematically related to other variables of interest. Unlike constant errors that are evenly distributed across all cases or random errors that occur randomly across all cases, systematic errors increase or decrease in unison with other variables included in the model being tested. Whereas constant errors have no effect on the relationships between key variables, and random errors reduce the strength of true relationships, systematic errors can result in relationships where none exist or can mask even strong relationships. The end result is that random errors can lead us to reject a predicted relationship when one really does exist (Type II error); but systematic errors can result in researchers accepting a predicted relationship even though it doesn’t in fact exist (Type I error) or rejecting one that does exist (Type II error). Although systematic errors can often be corrected when they are known, many are never fully acknowledged or understood.

The sources of systematic error are many. First, respondents may be more likely to inflate or deflate their responses to select questions based on their gender, race, income, education, or many other measures closely related to the dependent variable of interest. In a series of simulations (not shown) we found that if church attendance is systematically underreported or overreported based on gender, the persistent gender effect on religion can either dissipate or be greatly accentuated. A second source of systematic error is through a pattern of missing data. For surveys this often occurs because a group of respondents are more available or cooperative and as a result are more frequently included in the sample. For example, Darren Sherkat (2007) has recently argued that the non-response rate is higher for the most conservative Christians and this has resulted in surveys of political behavior overestimating support for liberal candidates. Thus, to the extent that the group of missing cases is related to other variables of interest, the results will be distorted.

Yet, a third source of systematic error, and one that we want to explore in greater depth, happens when a particular type of case or subgroup is not part of a dataset. This is especially common in population censuses. Like the missing cases in surveys, the concern is not simply undercounting, it is “differential undercounting” (Clogg, Massagli, and Eliason 1989). Is the undercount related to age, race, sex, and variables of interest? For population censuses, the concerns are partially political, but this differential undercounting also has important implications for theory testing. The implications can be illustrated with the Religious Congregations & Membership Study (RCMS). The RCMS approximates a census of congregational membership for every county, state, and urban area in the United States, and has proven valuable for multiple research projects (Beyerlein and Hipp 2005; Lee and Bartkowski 2004). Yet, even this major collection suffers from serious measurement errors, including differential undercounts that result in systematic errors.

The most obvious shortcoming of this collection is that it is not a complete census. The 149 groups included in the RCMS collection fall far short of the 2,600 groups listed in J. Gordon Melton’s (2002) *Encyclopedia of American Religions*. Because the vast majority of Melton’s groups are extremely small, however, the omission of most of the groups introduces little or no systematic error into the measure of religious adherence. For example, missing the 25 members of Mahasiddha Nyingmapa Center in Massachusetts will do little to change statistical relationships. The major exception, however, is the omission of the historically African American denominations such as African Methodist Episcopal Zion, Church of God in Christ, National Baptist Convention of America and others. With the majority of African American adherents missing from the measure

11Few findings have been more consistent over time and across regions than the effect of gender on religion (Miller and Stark 2002; Roth and Kroll 2007; Stark 2002). The BRS also finds a strong gender effect with 44% of the women reporting attending religious services at least weekly compared to only 32% for the men. Gender (female = 1) and religious service attendance are significantly correlated, with women attending with greater frequency (b = .140, p < .01). If we assume, however, that there is systematic underreporting by males and increase each male’s church attendance by 1, the correlation between gender and church attendance drops from significance (b = -.034). If, however, we assume systematic overreporting by males and decrease each male’s church attendance by 1, the correlation between gender and church attendance jumps to .301 (p < .01).
of religious involvement, this systematic error distorts the relationship this measure holds with other variables related to race.  

In an attempt to correct for these systematic measurement errors, Roger Finke and Christopher P. Scheitle (2005) computed correctives for the 2000 RCMS data that estimate adherent totals for the historically African American groups and other religious groups that are not provided in the original data. Using these data, Table 3 reports a series of bivariate correlations using rates computed from the original RCMS collection and adjusted rates using the correctives. The upper rows of the table report on bivariates for Alabama counties, where 26 percent of the population is African American, and the lower rows report on Iowa counties where the African American population is only 2 percent.

Table 3: Bivariate correlations with RCMS adherence rates adjusted and unadjusted for undercounts of African-American and ethnic congregations

<table>
<thead>
<tr>
<th></th>
<th>Alabama N=67</th>
<th>Iowa N=99</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unadjusted Adherence Rates</td>
<td>Adjusted Adherence Rates</td>
</tr>
<tr>
<td>% Below Poverty Line</td>
<td>-.64</td>
<td>.03</td>
</tr>
<tr>
<td>% Population Change, 90-00</td>
<td>.07</td>
<td>-.47</td>
</tr>
<tr>
<td>Larceny Rate</td>
<td>-.20</td>
<td>.03</td>
</tr>
<tr>
<td>% African-American</td>
<td>-.74</td>
<td>.04</td>
</tr>
</tbody>
</table>

Looking at the correlations for the Alabama counties we can see dramatic shifts once the correctives for systematic errors are entered. For example, the unadjusted adherence rates hold a strong negative correlation with the percentage below the poverty line ($r = -.64$), but the relationship fades away ($r = .03$) once corrections are made for the undercounted adherents. Hence, the adjusted rates correct for the systematic under reporting of African American adherents and prevent a Type I error. In the second row, however, the uncorrected adherence rates show no relationship to population change and the adjusted rates show a strong negative correlation ($r = -.47$). Here the adjusted rates unmask a true relationship and prevent us from rejecting a hypothesis that religious adherence is related to population change (Type II error). Each row of correlations for Alabama counties show how differential under reporting by race, a measure that is closely related to church adherence, results in systematic errors that lead to both Type I and Type II errors. Finally, as expected, the correlations for Iowa show that the correctives make little difference for a state with little systematic bias in the measure.

The RCMS measure of religious adherents offers an example of systematic measurement errors that are easy to recognize and relatively easy to correct. But most systematic measurement errors are more subtle and are seldom easy to correct after the data have been collected. Recognizing the dangers of measurement errors is far easier than avoiding the errors. To avoid many measurement errors we need to take a step back and review how measures are developed and evaluated. Below we review a few of the areas more evaluation is needed and new measures are required.

Reassessing Religion Measures

12For example, it will be of significant concern for those studying religion and crime (Alba et al. 1994; Evans et al. 1995; Heaton 2006; Bainbridge 1989; Liska et al. 1998).
Social scientific studies of religion strive to provide observable measures for each concept of interest. But the measurement of a concept can break down in multiple areas. Sometimes the concept is poorly defined and lacks the precision needed for evaluating what should be included or excluded. Other times the concept is clearly defined, but finding an observable measure is nearly impossible. Still other times we find a valid measure for only a portion of the population of interest. Perhaps the most common breakdown, however, is simply the design of the survey questions. Here we review several of the challenges, as well as the common errors made, when providing observable measures for abstract concepts.

Theoretical Clarity and Precision

The study of religion is plagued with vaguely defined concepts. Yet, conceptual clarity is the essential starting point for developing precise measures. The definitional boundaries of the concept should provide a clear guide for evaluating a measure. In other words, clear definitional boundaries provide criteria for evaluating if a measure is including and excluding the appropriate phenomena. Unfortunately, however, the process often works in reverse. Too often the measures used in a study serve to define the concepts rather than the concepts guiding the measures developed and selected. When studying religion, there is no shortage of vaguely defined concepts. Indeed, some of the most frequently used concepts are also the most vaguely defined.

In 1967, Larry Shiner complained about a “total lack of agreement as to what secularization is and how to measure it” (Shiner 1967: 207). He identified six “types of usage” of the concept and reviewed examples of how each type was applied in research. When summing up his findings, he noted that, “the appropriate conclusion to draw from the confusing connotations and the multitude of phenomena covered by the term ‘secularization’ would seem to be that we drop the word entirely (p. 219).” Sensing that this wouldn’t happen, however, he went on to suggest that researchers carefully state their “intended meaning and to stick to it.”

If Shiner could identify six “types of usage” of the concept of secularization in 1967, we can no doubt find scores more today. But if secularization is one of the most vaguely defined concepts, it is not alone when it comes to hazy conceptual boundaries. The concept of religion is often neatly divided into substantive and functional definitions; yet within these two categories many variations continue to exist (Christiano et al. 2002; McGuire 2008; Roberts 2004). Despite significant efforts to define the concept of socio-cultural tension, an idea which has been utilized extensively in the sociology of religion to differentiate between sect-like and church-like religious groups, the concept remains particularly difficult to measure (Bainbridge 1996; Stark and Bainbridge 1985; Stark and Finke 2000). Various indicators such as a religious group’s moral strictness, its social or theological conservatism, its level of exclusivity, its social class composition and its connection to one of several historical religious traditions have served as proxies for socio-cultural tension in the past (Bainbridge and Stark 1980; Stark and Finke 2000). Recent work on the production of social capital within religious groups underscores the conceptual ambiguity of yet another popular but controversial concept in the study of religion (see Paxton 1999). Researchers would do well to heed the advice of Shiner: we should carefully and clearly define our intended meaning and stick to it.

Connecting Theory and Measurement

Too often we have measures that are in search of a concept. Measures are often selected for their utility or availability rather than because they measure key theoretical concepts or address significant theoretical questions. In fact, measurement innovations are frequently independent of theoretical developments and often defy precise definitions for what they are measuring.

Some of the most heavily used indicators in religion research measure a medley of ill-defined concepts. For instance, from the groundbreaking survey work of Glock and Stark (1965) to the more recent efforts of the RELTRAD measure proposed by Steensland et al. (2000), social scientists have long attempted to classify denominations along meaningful spectrums. Everyone knows, however, that the final categories used for denominational affiliation are a mishmash of historical, theological, and social characteristics, obligations, and beliefs (Dougherty et al. 2007; Woodberry and Smith 1998; Kellstedt et al. 1996). As a result, the indicator is used as a proxy measure for a diverse array of concepts (Keysar and Kosmin 1995; Peek et al. 1991; Smith 1991). In our work alone, we have used it as a proxy for tension with the dominant culture, organizational strictness and conservative theological and moral beliefs. Despite the power and intuitive appeal of this measure,

13A second alternative that he offered was for researchers to agree on the term as a general designation or large scale concept covering certain subsumed aspects of religious change (Shiner 1967: 219).
it lacks the precision needed because it conflates so many concepts of interest. For instance, knowing that adherents who fall into the evangelical category give more money, doesn’t fully address the question of why they do so. A well-developed measure based on clear theoretical concepts should provide some explanation for the relationship between variables of interest.

The frequently used measure on biblical literalism is even more problematic. Used as a proxy for the measure of religious orthodoxy, fundamentalism and a wide range of other concepts, the measure is limited to a Christian sample and the response categories fail to measure important differences between respondents. According to both the GSS (2004) and NES (2004), one-third of Americans (33.2% and 37.1% respectively) believe that the Bible should be interpreted literally. More than 80 percent believe that the Bible is the Word of God, even if they do not believe that it should be taken literally. It would seem, then, that a majority of Americans hold uniform views on biblical authority. Yet, the public and often rancorous denominational battles that have been waged over the issue of biblical interpretation (see Ammerman 1995) belie the fact that important disagreements about the Bible divide many Americans. So, why do survey data consistently fail to pick up on these fault lines? The response categories offered for questions such as this one are not specific enough to pick up on meaningful differences. Americans’ beliefs about the Bible are more nuanced than responses to this question suggest, and so it would be helpful if survey questions were developed to tap these views (Bartkowski 1996; Woodberry and Smith 1998).

The fit between measures and concepts can also be reduced due to religious changes over time and the rise of new theoretical or substantive questions. In recent years, much attention has been paid to the sudden rise of Americans reporting no religious preference — a group often referred to simply as the religious “nones.” According to GSS data, the percentage of “nones” in the U.S. doubled during the 1990s, increasing from 7 percent in 1991 to 14 percent by 2000 (Hout and Fischer 2002). But the category of “nones” poses many questions that can’t be addressed with the current measure. Hout and Fischer (2002) find that a significant number of “nones” pray and hold conventional beliefs such as belief in God, heaven, and hell. Moreover, researchers using BRS data found that 6 percent of individuals claiming no religious preference actually report affiliation with a denomination or congregation when asked (Dougherty et al. 2007). And the Pew Forum’s 2008 U.S. Religious Landscape Survey of more than 35,000 respondents found that most who report no religious preference are neither agnostics nor atheists (see Table 4). Hout and Fischer (2002) contend that politically moderate and liberal Christians are increasingly likely to claim no religion in response to some Christian groups’ alignment with conservative politics. Others question this claim (Marwell and Demerath 2003). The entire discussion suggests that the current measure of religious preference, which has changed little over the years, fails to adequately measure contemporary variations in Americans’ religious identity.

Table 4: Refining Response Categories

<table>
<thead>
<tr>
<th></th>
<th>GSS</th>
<th>NES</th>
<th>Pew Forum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protestant</td>
<td>53.0</td>
<td>56.1</td>
<td>51.3</td>
</tr>
<tr>
<td>Catholic</td>
<td>23.4</td>
<td>24.9</td>
<td>23.9</td>
</tr>
<tr>
<td>Jewish</td>
<td>2.0</td>
<td>2.9</td>
<td>1.7</td>
</tr>
<tr>
<td>Other</td>
<td>7.3</td>
<td>.9</td>
<td>7.0</td>
</tr>
<tr>
<td>No Preference: Agnostic</td>
<td>2.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Preference: Atheist</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Preference: Total</td>
<td>14.4</td>
<td>15.1</td>
<td>16.1</td>
</tr>
</tbody>
</table>

Measures for All

Moving beyond the connection between theory and measurement, the indicators used can also break down because they fail to serve as measures for the entire population of interest. As reviewed earlier, the most serious measurement errors are those that are systematically related to another variable of interest. Earlier we used the example of indicators not being effective measures for all races, but one of the greatest challenges for

14Attempts to classify respondents based on their religious identity, face similar challenges (see Alwin et al. 2006).
research on religion is asking questions that are effective measures for all major religions and cultures. With an increasing number of cross-national studies and an increasing religious pluralism in the U.S., this is a challenge facing anyone doing survey research.

Cross-cultural measures can fail at many levels. One of the most obvious and, yet overlooked, is translation. Ensuring uniform meaning across respondents is always a challenge, but when writing questions for multiple languages with linguistic variations within those languages the sources for error are multiplied. Tom Smith (2004:446), the past secretariat of the International Social Survey Programme (ISSP), recently wrote that, “no aspect of cross-national survey research has been less subjected to systematic, empirical investigation than translation.”

Similar challenges hold when conducting research across major religious groups. Appropriate questions on belief and practice will vary from one group to the next. A common solution is to simply ask Muslims, Christians, Hindus, Jews, and others different questions — questions that are specific to their religion. For instance, on the 2005 World Values Survey (WVS) respondents in Islamic societies were asked about the frequency with which they prayed, whereas respondents in other countries were asked about the frequency of their participation in religious services. In addition, WVS respondents in non-Christian societies were asked about the public role of religious authorities rather than the public role of churches. The obvious problem with these adaptations is that the results from each group can no longer be compared to the results of other major religions. Adapted questions may be similar, but they are not equivalent. Needless to say, the preferred option is to develop questions that can be used across religions and cultures; questions that can measure a more abstract concept without relying on terms, rituals or texts that are specific to only one religion.15

Sometimes the changes needed are as simple as using more inclusive language. Asking respondents about the sacred text of their religion, as opposed to the Bible, is an example. When the BRS asked the question, “Outside of attending religious services, about how often do you read the Bible, Qur’an, Torah or other sacred book?,” only 24 percent of the respondents indicated that they never read such books. By contrast, when the 1998 GSS and 2000 NES asked about “Bible” reading, 42 percent of the GSS and 38 percent of the NES reported never reading the Bible. Revising survey questions from being culture and religion specific to being cross-cultural often requires much more than a change in wording. Once again, the starting point is returning to the abstract concept being measured. What are we trying to measure and how is it defined? For instance, are we really trying to measure concepts such as biblical literalism and denominational affiliation? Or are we trying to measure an exclusiveness of religious beliefs, behavioral requirements, certainty of beliefs or density of religious networks?

Examples of promising measures with cross-cultural utility are those measuring the images of God. Because the practice of all major world religions is centered on beliefs in a deity, the questions can apply uniformly even when using cross-national samples. Andrew Greeley was a pioneer in this area, developing a series of items measuring conceptions of God that have appeared in the GSS and ISSP. Using these items, Greeley (1988, 1989, 1991, 1993 and 1995) found significant correlations between images of God and political and social views. More recently, Christopher Bader and Paul Froese (2005, 2007) have built on this work theoretically and methodologically. Theoretically they helped to expand the implications that images of God have for many behaviors and beliefs. Methodologically they added a more refined set of measures on images of God to the most recent BRS. Work in this area is just beginning and requires far more evaluation, but the initial results for developing cross-national measures are promising.

Attention to Design

Much of our attention has focused on the connection between abstract concepts and the measures used. Yet, many measurement errors are the result of far more mundane problems related to the cognitive psychology of taking surveys. These seemingly mundane design problems can have significant consequences. We offer one example, but many others could be garnered.

The American Congregational Giving Study (ACGS) conducted in 1993 (Hoge et al. 1996) and the National Congregations Study (NCS) conducted in 1998 (Chaves 2004) are two highly regarded congregational surveys. The two surveys had different research objectives, but each asked a series of comparable questions

15Tom Smith (2004: 445) distinguishes between emic and etic questions. Etic questions have “a shared meaning and equivalence across cultures, and emic questions are items of relevance to some subset of the cultures under study.”
about programs and groups within congregations. Their findings in this area were strikingly different. For instance, 0 36 percent of the congregations in the NCS reported having a women’s group, compared to 92 percent of the congregations in the ACGS. The NCS found 5 percent of congregations reporting single’s groups, compared to 33 percent in the ACGS. In addition, the NCS found 10 percent reporting musical groups other than the choir, while the ACGS found that 72 percent of the congregations reported these groups. Not all of the questions can be compared, but the trend is clear: the ACGS was finding far more groups and programs in the local congregation.

What explains this disparity? Although some differences may be the result of sampling or other design issues, the differences are largely the result of measurement design and the prompts used to encourage recall. The ACGS (1996) asks respondents about the existence of many specific types of groups. In other words, respondents are specifically asked if their congregation has a prayer group, a woman's group, a Bible study group and so on. In contrast, the NCS asks respondents to name the purposes for which congregational groups regularly meet, leaving it up to the respondent to recall specific types of groups. When respondents are asked to recall all of the groups and programs in their congregation without prompts, they name far fewer than when they are given a checklist.

The importance of prompts in shaping findings is further confirmed by other recent studies of congregations and volunteers. When Ram Cnaan and his team of interviewers asked clergy in Philadelphia about specific types of programs, they found that the number of programs increased sharply as they used prompts to facilitate recall (Cnaan et al. 2002). Likewise a recent article in the Nonprofit and Voluntary Sector Quarterly finds that longer and more detailed prompts lead respondents to recall higher levels of volunteering and donating (Rooney et al. 2004), even when controlling for other design and sampling differences. Several measurement issues remain in this area. The most obvious is determining the appropriate balance between prompting for recall and actually shaping recall. Cnaan et al. (2002) have found that a lack of prompts results in under reporting; but does the extensive use of prompts result in inflated reporting? Second, do such measurement errors result in systematic errors, random errors or only constant errors? For the purpose of this paper, the most significant point is that the design of the measure and the prompts used clearly shape the results.

Setting an Agenda

Measurement error is ubiquitous in the social sciences and few, if any, would suggest that it can be eliminated. But steps can be taken to both reduce the errors and to better identify the source of the errors. Building on the issues raised in this paper, we try to identify practical steps for assessing existing religion measures and developing new ones. We recognize that they are far from comprehensive, but we offer this list as initial steps for improving religion metrics.

1. Conceptual clarity
An essential starting point is working with clearly defined concepts and measurement models. Conceptual definitions should draw boundaries for what is included and excluded within the concept we are attempting to measure. Just as it is hard to hit a bull’s eye when the target is unknown, it is difficult to measure a concept with any specificity when the conceptual boundaries are ill-defined. We noted earlier the lack of clarity with some of our most basic concepts such as "religion" and "secularization," and many other topics are studied at length without being adequately defined. For some measures this might be the result of scholars’ commitment to a social constructionist approach to the study of religion, but for most it is simply a lack of clarity. Therefore, we advocate first developing clear and specific definitions for the concepts that are to be studied. If there is disagreement among scholars over the best way to define a concept, as there often is for such important concepts as "religion" and "secularization," researchers must delineate clearly what definition they are using and how concepts are operationalized.

2. Increased attentiveness to systematic errors
Earlier we demonstrated that systematic errors are the most dangerous for theory testing because they can result in relationships where none actually exist or can mask even strong relationships that do exist. Yet, for most key religion measures we have few studies that examine the way measurement errors systematically vary by demographic and social groups of interest. This is an area that deserves attention. For instance, in the case of church attendance there are multiple studies attempting to measure the extent of reporting error (e.g., how many people actually attend church each week), yet we know little about how this error varies by gender, race, region,
or other key correlates of religion. For example, do respondents from more religiously active regions of the U.S. feel pressure to report church attendance?

Gaining an understanding of the effects that systematic errors are likely to have on key religion measures such as church attendance, frequency of religious practices and religious affiliation will require new studies that make it possible for researchers to compare respondents’ self-reported behaviors with their actual behaviors. Indeed, measurement studies in other substantive areas, such as delinquency, have clearly shown that underreporting and overreporting varies systematically by sex and race (Hindelang, Hirschi, and Weis 1981). Similar studies should be conducted for key religion measures.

With attention typically drawn to explained variance and the significance of the coefficients in multivariate regression models, we often fail to assess if systematic measurement errors are suppressing or inflating reported relationships.

3. Using experimental designs to develop and pre-test new measures.

We have identified a host of current and potential problems with religion measures. Some are measuring a medley of concepts, others are poorly worded or translated. Some are a poor fit with current theoretical and substantive concerns, others are limited to a subsample of our population of interest (e.g., Protestant Christians). The challenge for survey methodologists is identifying and remedying these problems before the measures are placed on the final survey. But identifying the problems early requires highly effective pre-test procedures. One of the most promising is combining experimental designs with survey pre-tests.

Survey researchers often treat the pre-test as a quick “dress rehearsal” before the big event. Even highly regarded survey methodologists have suggested that after 25 or more cases most problems with the instrument can be found (Sheatsley 1983; Sudman 1983). But a growing number of researchers are now recognizing that developing metrics that effectively measure specific concepts requires more than the traditional pre-test. This is especially true in the area of religion, where finding a shared vocabulary is difficult. Increasingly, focus groups are used prior to writing an instrument and various forms of interviews are used with respondents during and after the pre-test (Presser et al. 2004). Each of these techniques has proven a valuable supplement to existing pre-tests. Our attention is focused on the use of experimental design in pre-tests. We explain how such designs can be highly effective in evaluating measures.

The obvious advantage of experimental design for any research is that only one variable is changed while others are held constant, making it easier for the researcher to identify the source of an effect. In the same way, experimental design allows researchers to compare two questions that are identical with the lone exception of question wording, question ordering, response categories or the instructions given (Fowler 2004). Pre-test respondents are randomly assigned to either version 1 or version 2 of a particular measure, and then their responses to those measures are analyzed and compared. The power of this design can be illustrated with a simple classroom example.

Students at Penn State University were each given a computer administered survey in the Spring of 2001. For selected questions the software divided the students into two groups based on the date of their birthday (an odd or even day in the month). Group A was asked if the Mississippi River is longer or shorter than 500 miles. Group B was asked if the Mississippi River is longer or shorter than 3,000 miles. Immediately following this question, both groups were asked the same question: How long is the Mississippi River? When the preceding question asked if the Mississippi River was longer or shorter than 500 miles, the students estimated the river was less than 900 miles long. By contrast, students who were asked if the Mississippi River was longer or shorter than 3,000 miles estimated that the river was over 2,000 miles in length.16 Thus, the experimental design allowed us to isolate the effects of question ordering and quantify the differences between the two versions.

This classroom example illustrates how experimental design, especially when combined with computer administered and internet based testing procedures, can be used to aid researchers in pinpointing the source of measurement error before questions ever appear on an actual survey. Moreover, it also points to the limits of

---

16Eighty-eight Penn State students took the survey during the Spring Semester of 2001. For Group A the mean and median were 2,415 and 2,000. For Group B, the mean and median were 870 and 700. When conducted with 75 students at Purdue University in 1998 the differences between the means were slightly reduced, but remained over 1,000 miles apart.
respondents’ ability to evaluate questions and their own responses to those questions. Even after seeing the results, many students claimed that the preceding question had no influence on their estimate of the length of the Mississippi River. Finally, this example offers a glimpse at the advantages of using computer administered pre-tests — advantages that we explore more fully below.

4. Using the internet for pre-testing new measures.
When administering a pre-test, especially one with an experimental design, using a computer-assisted data collection over the internet offers many advantages. First, for experimental designs an obvious advantage is that the software can easily assign respondents to one of two or more groups. This assignment can be purely random or it can be based on more complex selection procedures (e.g., ensuring that each group has adequate representation by race, age or other variables of interest). Second, this mode of collection is highly flexible. Once the pre-test is underway, a question(s) can be revised or even replaced if it becomes clear that there are serious problems. Or, the percentage of respondents going into one group or another can be shifted on the fly. Third, this pre-test mode is less limited by space and location, a major advantage for cross-national pre-tests. The surveys can readily be administered around the globe, yet the data continues to flow to a single location. Fourth, potential new measures can be quickly introduced and tested. By nearly eliminating the time needed for setting up interviews and coordinating a site for administering pre-tests, new measures can be tested with little delay. Plus, the costs of online collection are low.

We recognize, of course, that the sample of respondents for such pre-tests will not be a representative sample of the larger population. For cross-national research, in particular, many will not have ready access to a computer. But even here this mode of pre-test can have advantages over current pre-tests relying on convenience samples in a single location. We also should acknowledge that using the Internet to collect pre-tests will not eliminate the need for in-person interviews and focus groups. These techniques will continue to provide information not available from online pre-tests. The online method does, however, reduce many of the hurdles for introducing and pre-testing new measures. With these hurdles removed, more new measures can be introduced and tested.

5. Systematic evaluation of existing measures.
Needless to say, the evaluation of measures doesn’t end after the data are collected. Many sophisticated methods are used in assessing reliability and for developing measurement models and indexes. Here we want to highlight more general concerns about the measures used and the evaluations they should receive. Just as we begin the measurement process by asking what we want to measure, we begin the evaluative process by asking whether it was actually measured.

First, to the extent that the concept is clearly defined (see #1 above), evaluating the face validity of the measures is improved. When the boundaries of the concept are clear, the boundaries of the measure will also be clear. For instance, defining religious affiliation as the particular denomination or religious group of which someone is currently a member makes it rather easy to assess the face validity of affiliation measures. Granted, this definition of affiliation may be too narrow to include many forms of religious affiliation that researchers would be interested in studying. Still, it provides an example of the way that clear and specific conceptual definitions increase our ability to determine a measure’s validity.

It is also possible to utilize other closely related concepts to test the validity of measures. For example, researchers can evaluate the statistical relationship between an indicator that they are using for a concept and other known measures of the same or similar concepts to test for convergent validity: whether the separate indicators actually measure the same thing. If the indicators are closely related, then researchers can be more confident in the validity of their measure. If, however, the indicators are divergent or poorly correlated this suggests that the measure may not be valid. In a recent paper examining various religious identity measures, Alwin et al., (2006) explore the statistical relationship between denomination-based religious identity measures and more subjective self-reported religious identity labels. Interestingly, their analyses reveal that these different ways of measuring religious identity actually tap two different aspects of religious identity. As a result, they suggest that a more accurate measure of identity might be constructed using both measures. We argue that religion researchers should carry out similar tests on measures of concepts such as religion, secularization and socio-cultural tension.

Researchers can also evaluate the relationship between a measure and the outcomes to which it is theoretically related. This type of analysis allows researchers to test for construct or predictive validity: the extent to which an item predicts outcomes that it is believed to predict (e.g., the extent to which conservative
religious identity predicts biblical literalism). If the relationship between a measure and an expected outcome is significant, then researchers can be more confident in the validity of their measure. If, however, the measure is not related to an outcome or is related in a way that is contradictory to that which is expected, then the validity of the measure may be called into question. Once again, Alwin et al. (2006) provide an example of how this type of analysis contributes to our understanding of a measure’s validity. Their study on measurement of religious identity utilized outcome measures that were believed to be highly correlated with religious identity (e.g., biblical literalism, belief in an afterlife, frequency of prayer, and church attendance) in order to test for predictive validity (p. 539). Their analyses reveal that different measures of religious identity are related to these outcome measures differently, suggesting that they tap different underlying concepts. Again, we argue that religious researchers should utilize similar tests to examine the validity of measures. Only after such rigorous tests have been conducted can researchers be confident that survey items actually measure what they are intended to measure.

Finally, as highlighted in #2 above, we need to continue sorting out the type and level of measurement error. We especially need to focus on understanding how systematic errors are introduced and what effects they have on hypothesis testing. This will likely require the completion of studies designed to observe the difference between actual and observed religious behavior as well as the development of improved statistical methods for detecting and accounting for systematic errors in collected data.

**Conclusion**

Stanley Payne’s (1951) classic text on survey design was entitled *The Art of Asking Questions*. Here we have stressed the science of assessing what these questions measure. The purpose of this paper is to highlight the importance of refining religious metrics. Even with sophisticated statistics and clearly stated concepts, we can never confidently test even the most basic theoretical models unless we develop adequate measures. We have tried to raise a few of the most central issues for measurement design. First, we illustrated how measurement errors occur and evaluate how these errors distort our research. We have argued that when testing relationships constant errors are typically less harmful than they might appear, random errors attenuate relationships and systematic error are the most egregious, potentially masking even strong relationships or finding relationships where none exist. Second, we offered a handful of examples to illustrate sources and types of measurement error in past research. We have shown how errors arise from a lack of theoretical clarity, weak or limited measures and faulty research designs. Finally, we offered a partial agenda for refining religion metrics. Here we have tried to outline some initial steps that can be taken to develop new measures as well as improve existing ones. Although many of our examples have been drawn from surveys, the points raised are important for all social scientific research. Regardless of the research design, measurement matters.
References


