PSJ Policy Studies Journal

Policy Studies Journal, Vol. 0, No. 0, 2018

Measuring Complex State Policies: Pitfalls and Considerations, with an Application to Race and Welfare Policy

Eric Plutzer, Michael B. Berkman, James Honaker, Christopher Ojeda, and Anne Whitesell

Welfare policy is multidimensional because of the political compromises, competing goals, and federalist structure underpinning it. This complexity has hindered measurement and, therefore, the comparability of research on race and welfare policy. This paper describes a measurement strategy that is transparent, replicable, and attuned to matching the assumptions of statistical models to the policy process. We demonstrate that this strategy leads to more nuanced conclusions regarding the relationship between minority caseloads and the flexibility of state welfare policies. The strategy and recommendations are adaptable to research agendas that scholars bring to the comparative study of welfare in the U.S. states, countries, or other units—and to other complex policies enacted in federal systems.

KEY WORDS: policy measurement, welfare policy, race

由于政治妥协、政策目标相互冲突，以及给予其支持的联邦制的结构特征等上述现象的存在，福利政策具有多维复杂性。这种复杂性阻碍了对福利政策的定量测量，也因此阻碍了种族和福利政策方面的比较研究。本文描述了一种清楚易懂的、可复制的测量策略，并且它符合政策过程统计模型的假设。本文证明，根据这一策略，我们可以对少数族裔数量与国家福利政策灵活性之间的关系得出更细致入微的结论。该策略和建议适用于研究学者对美国各州、以及各国家及地区的福利政策进行的比较研究，同时也适用于对联邦制内所实施的其他复杂政策的研究。

This special issue encompasses recent research on immigration policy and racialized policies, two domains that often overlap and which pose strikingly similar challenges to empirical researchers. Along with incarceration and segregated schooling, cash assistance and related welfare programs are among the most racialized policies in the United States. In that light, it is not surprising that comparative state welfare studies have contributed important insights into political parties, elections, diffusion, social control, immigration, and insurgency (e.g., Cnudde & McGrane, 1968; Dawson & Robinson, 1963; Dye, 1984; Filindra, 2013; Fording, 1997; Fording & Berry, 2007; Hill & Leighley, 1992; Hill, Leighley, & Hinton-Andersson, 1995; Hofferbert, 1966; Key, 1949; Lockard, 1959; Piven & Cloward, 1993; Plotnick & Winters, 1985, 1990). Nearly all of this research recognizes that state welfare policy is doubly racialized in that policies affect racial and ethnic groups

doi: 10.1111/psj.12307
© 2018 Policy Studies Organization

differently and racialized cultures and ideologies shape policy (especially Gilens, 1999; Lieberman, 2001; Manza, 2000; Orr, 1976; Schneider & Ingram, 1997; Schram, 2005; Soss, Fording, & Schram, 2011; Wright, 1977).

This large body of literature, however, raises the question of how state welfare policy should be measured. The challenge of measuring a complex policy is the subject of many papers in this special issue (Bjerre, Römer, & Zobel, 2018; Goodman, 2018; Monogan, 2018; Reich, 2018). These papers grapple with more general questions of comparative policy measurement and most especially questions concerning whether specific aspects of policy should be combined into summary scores that measure legislative output. In this paper, we introduce a seven-step approach that allows scholars to answer this question in a way disciplined by theory, prior research, and best practices from the science of measurement. Our empirical results show how much substantive conclusions can vary if the dependent variable is defined too narrowly or if components are added to an index without the discipline of strong theory. We develop recommendations using the specific example of comparative welfare policy among the 50 U.S. states, and we begin by providing some background on welfare policy and why it poses particularly thorny challenges to policy measurement.

Prior to the 1996 Welfare Reform, research coalesced around common measures of benefit generosity (e.g., Tweedie, 1994), state tax effort (Jennings, 1979; Pacheco, 2013), and redistribution (Plotnick & Winters, 1990). In contrast, the complexity of Clinton era reforms led to a correspondingly diverse set of measurement approaches. This diversity renders many major studies incomparable to one another, makes it difficult to reconcile contradictory findings, and limits the cumulative growth in our understanding. Moreover, the salience of race in comparative welfare research can have unrecognized consequences for measuring policy—when that occurs, policies may be measured in ways that tip the scales so that some empirical conclusions are more likely than others.

The contributions of this paper are threefold. First, we highlight common practices in the measurement of social policy generally that are often theoretically and statistically inappropriate. Second, we describe a seven-step approach to complex policies that is transparent, replicable, and yet adaptable to the many research agendas that scholars bring to the comparative study of welfare, immigration, environmental protection, the current opioid epidemic, and other complex policy domains. Third, we demonstrate the usefulness of our approach by applying it to the racialization of welfare policy.

Then analyses show how this approach yields new findings that contradict those of the most influential works in the field. The Racial Classification Model, and work in the constructivist tradition more generally, understand policies through the lens of racial stereotypes. Previous scholars have assumed that these stereotypes, when activated, will contribute to policies that impose a greater burden on the target population. While this assumption may be true on average, exceptions can be very important. The defining feature of welfare reform is "work" and yet considerable scholarship has overlooked most of the relevant work-related policies. Analyses guided by this seven-step approach to measurement show that racialization operates in a counterintuitive way in how states define "work."

As background, the next section explains how the 1996 welfare reform dramatically expanded the number of consequential moving parts of state welfare policies and then reviews how leading scholars translated this complexity into dependent and independent variables. The following section highlights avoidable dangers in the most commonly used measurement strategies and shows how these can lead to misleading conclusions about how race influences policy formation. This section is followed by an elaboration of a third approach that rests on stronger methodological ground and illustrates it by describing an original data set measuring state TANF policy from 1997 to 2016.

Measuring State Welfare Policy: One Data Source Used Many Ways

Spurred by the widespread use of state welfare waivers, The Personal Responsibility and Work Opportunity Reconciliation Act of 1996 replaced Aid to Families with Dependent Children with Temporary Assistance for Needy Families (TANF), which presented states with unprecedented choices. For example, eligibility was no longer determined primarily by a means test, as states had to consider recipients' efforts to find work; the type and amount of work; an array of time limits and penalties; and rules intended to shape parenting behavior, increase child support from absentee fathers, and much more. The complexity of state laws created a pressing need to develop new measures of state welfare policy.

Almost all TANF researchers rely on the data contained in the Welfare Rules Database (Urban Institute, 2017). Each year since passage of PWORA in 1996, the Urban Institute's research staff read every state's caseworker manual and apply codes based on consistent descriptors; they then send a Databook summarizing the codes to state welfare administrators for verification. Researchers have a nearly infinite number of options in using these codes, which apply to every stage of the welfare experience in each state and are presented on the Urban Institute web page from the perspective of someone seeking or receiving cash assistance. Rules are coded as either present or absent, or by textual descriptions that can be classified.

Scholars have utilized the Welfare Rules Database in various ways. Many measure the presence or level of a small number of policy elements (e.g., Avery & Peffley, 2005; Monnat, 2010; Soss, Schram, Vartanian, & O'Brien, 2001). Others select a larger number of policy rules and create composite scales, often using data reduction techniques such as item analysis (e.g., De Jong, Graefe, Irving, & St. Pierre, 2006), cluster analysis (McKernan, Bernstein, & Fender, 2005), or Item Response Theory (IRT) models (Berkman, Honaker, Ojeda, & Plutzer, 2013b); others combine rules in a more *ad hoc* manner (Fellowes & Rowe, 2004). We argue that each of these comes with serious, sometimes unrecognized, limitations.

Researchers selecting a small number of policy elements must choose from among the more than 550 rules that specify who has to work, what counts as work, how much work is required, how long one may receive cash assistance, who can waive the lifetime limit, whether recipients need to immunize their children, and which violations trigger small sanctions, more substantial penalties, or complete loss of eligibility (to list just some of the options).

The most influential scholarship employing this approach is that of Soss and colleagues (2001, p. 380) who looked at four "get tough policy choices" that received the "lion's share of media attention:"

1. Demanding work from recipients *earlier* than the federal requirement of 24 months.
2. Adopting a lifetime eligibility cut-off shorter than the federal limit of 60 months.
3. Adopting a family cap, prohibiting additional funds for additional children.
4. The severity of punishments for rule infractions.

Of course, there may be important and valid reasons to focus on a small number of policy elements. For example, in developing a dependent variable that captures state immigration policies, Monogan (2018) reveals distinctly different results when policies are aggregated to policy subareas using a relatively small number of variables, rather than aggregated across policy subareas. Thus, the danger is not in the focus *per se*, but in the temptation to draw overly broad conclusions about "welfare policy" more generally. As we show below, this is especially consequential in research on policy racialization: some of the items excluded by Soss et al. (2011) evidence no linkage with race, while others show both positive and negative associations with the racial composition of caseloads.

The second approach, of combining rules to create composite scales, only partially addresses the challenge of selection. Fellowes and Rowe (2004) chose a set of 28 policies to create their *eligibility index*, and 12 rules to create their *flexibility index*. Other scholars have replicated these two scales (as best they could) and employed them in their own work (Kim & Fording, 2010; Reingold & Smith, 2012).

The use of data reduction models to construct composite scales would seem to solve the dilemma of which rules to include by estimating a "kitchen sink" model. In practice, however, scholars have examined only a fraction of the 550+ codes available for each state in the Welfare Rules Database. In the most ambitious data reduction effort to date, De Jong and colleagues (2006) use principles of Classical Test Theory to identify an initial set of 78 rules as candidates for scale inclusion. They then use factor analysis to *purify* their scales by dropping 35 rules that did not have a sufficiently high factor loading. Like others, they made conscious choices to exclude some policies from consideration, sometimes describing their inclusion and exclusion rules, but not in sufficient detail to enable replication.

Data reduction faces a second, more formidable, challenge—selecting a model that is a close analog to the data generating processes in the world of politics and policy. For example, Graefe, De Jong, and Irving (2006, p. 819) explain that their factor analysis assumes that "the new and increasingly stringent rules for reforming welfare eligibility" reflect a single underlying dimension: "public support for the poor." From this perspective, liberal states would have generous benefits, flexible work requirements, and milder sanctions than conservative states. In statistical terms, a single latent variable could account for the variance in a large number of policy elements. This assumption turned out to be false, as they could not reduce their 15 initial subscales to a single dimension.

Data Reduction Models Rarely Match Theory or the Policymaking Process

Data reduction models such as factor analysis and IRT are popular. While useful in many cases, a close examination of these models' assumptions reveals that they are often inappropriate, because they assume that one or more latent variables create spurious correlations among measured policy elements. Such *reflective models* were designed specifically to infer attitudes, beliefs, motivations, and other cognitive factors that account for observable outcomes. Figure 1 illustrates how a reflective model might apply to the measurement of welfare policy.

In this view, the latent variable is some aspect of policymaker ideology, such as left–right orientation, views about the state's role vis-à-vis the poor, or their racial ideology that reflects stereotypes and implicit racism. If a state has a high level of this latent variable, this *causes* it to adopt strict policy elements (e.g., stiff sanctions, short time limits). The strength of these causes is *reflected* in factor loadings (λ_1 to λ_4), but of course with some error (e_1 to e_4) that is presumed to be random and uncorrelated with everything else in the model (IRT models have more parameters, for the "difficulty" of each rule being adopted, but rely on the same assumptions).

When researchers estimate these models, some policy elements will not fit well as indicated by a small (or negative) factor loading (or a flat item discrimination curve if estimated by IRT). Sometimes several policy elements that do not "fit" will form their own factor, but sometimes will appear unique. For example, in the model depicted in Figure 1, suppose that the factor loadings for the first three elements are large and positive, but the loading for work requirements, λ_4, is small or negative.

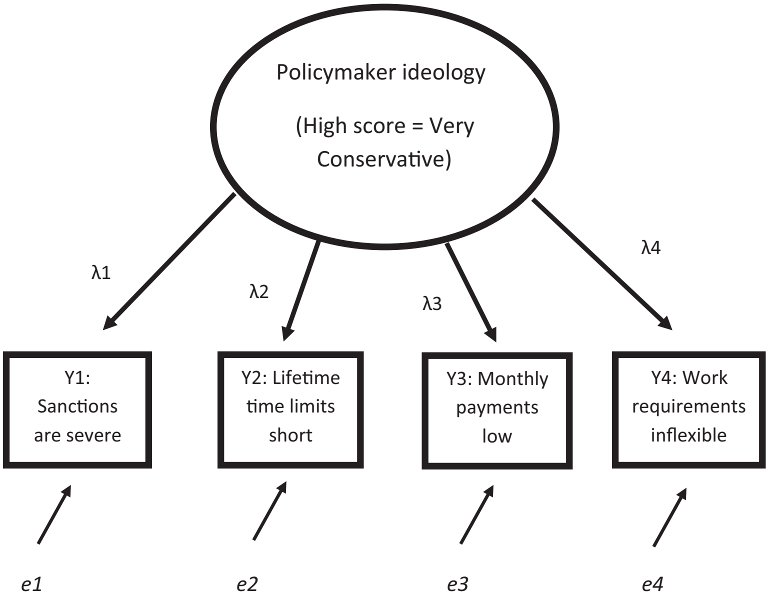
Policymaker ideology
(High score = Very Conservative)
λ1
λ2
λ3
λ4
Y1: Sanctions are severe
Y2: Lifetime time limits short
Y3: Monthly payments low
Y4: Work requirements inflexible
e1
e2
e3
e4

Figure 1. Common Factor Model Applied to Welfare Policy.

The nearly universal practice, *scale purification*, is to exclude this policy element from the resulting measure of state policy. This decision is often reinforced when a measure of internal consistency—such as Cronbach's α—increases when the offending rule is dropped from the scale. But dropping a rule is the correct decision if and only if the proposed model—in this case, that policymaker ideology is the dominant determinant of policy adoption—reflects how the world actually operates and if the main purpose of the measurement exercise is to understand how policy is formed.

There are two common circumstances when these assumptions are unwarranted. The first is when the purpose of the research does not concern policy generation. Suppose the goal is to develop a measure of policy severity from the recipient's perspective (the oval is labeled "policy severity" rather than "policymaker ideology"). Analysts might focus on severity if, for example, state welfare policy is the independent variable and mental or physical health, food insecurity, or civic engagement of the recipients is the dependent variable (e.g., Mettler & Soss, 2004). From the perspective of the recipient, inflexible work requirements (Y_4) are severe and may lead to stressful and adverse outcomes. Critically, this experience of stress does not go away simply because λ_4 is zero or negative. The negative loading only implies that states with severe sanctions do not consistently have inflexible work requirements (as shown empirically below). Scale purification reduces validity because it requires scholars to ignore a real burden experienced by millions of welfare recipients. Rather, researchers should tailor measurement to the research purpose at hand, and not be driven by the typical exclusion features built into data reduction and scaling methods.

A rarely appreciated, empirical implication of a client-centered model is that the causal arrows must flow upward—the rules cause burdens to be severe, not vice versa. Therefore, a residual attaches to the latent variable and not the indicators. Widely recognized in fields such as sociology, psychology, and marketing (Bollen & Lennox, 1991; Diamantopoulos, Riefler, & Roth, 2008) *formative* measurement models should not be estimated by the usual factor or IRT approaches.

The second common circumstance when reflective model assumptions do not apply is when policymaking reflects compromises and trade-offs. When this is the case, the manifest items (the boxes in Figure 1) will not necessarily be positively correlated with one another even if the rules reflect the latent variable. Ojeda, Whitesell, Berkman, and Plutzer (2017) argue that TANF saddles each state with an accountability regime that places its block grant at risk under some circumstances. These risks, in turn, force policymakers to forego ideologically motivated goals to make it easier for welfare recipients to be classified as "working." They show that these trade-offs become especially salient when racial stereotypes are activated. In these circumstances, conventional data reduction and scaling methods can lead to erroneous conclusions about policy formation.

In light of policy trade-offs, the failure to reduce 15 subscales to a single underlying dimension of policy stringency (De Jong et al., 2006), or to successfully reduce just four "get tough" policies to a single dimension (Soss et al., 2001) is unsurprising.

A similar problem occurs whenever policies are complementary and, therefore, negatively correlated. Consider the penalties that states can impose when a recipient fails to meet their work requirement. Under TANF, a small number of states close the case, even for a "first offense." Others impose a partial reduction in benefits, which is restored when the requirement is met. No state has both rules, and the negative correlation that this implies can lead to two poor research decisions—to exclude one item because it "does not scale" or (worse) to include it and accept that it gets a negative weight in the construction of scale scores. The former loses information, but the latter would incorrectly code case closure as contributing to generosity! Step 4, below, addresses this in greater detail and provides suggestions for coding complementary policies. A similar problem arises with the family cap, a frequently studied rule that limits monthly benefits as family size increases. An important element in family cap policy is the number of exemptions granted (such as to recipients whose pregnancies are the result of incest or rape), yet these exemptions are typically excluded from empirical analyses because states with values of zero are a heterogeneous mix of the most liberal policies (no cap) and the most conservative policies (cap, with zero exceptions). Whitesell (2017) addresses this by combining these two rules into a summary measure.

To sum up, the most widely used data reduction models—Classic Test Theory, IRT, factor analysis, and cluster analysis—all assume that the data-generating process is captured by a reflective latent variable model (rules are caused by the latent dimension), with no policy trade-offs or complementary rules. Treating complex policy elements as if they are questions in a personality test will lead researchers to exclude some rules altogether and combine others in ways that compromise validity.

A Third Way: A Seven-Step Approach to Measuring Complex Public Policies

This section details a seven-step approach to policy measurement. This approach relies on scholars' substantive understanding of the policy-generation process, while simultaneously limiting scholars' subjective judgments about inclusion and exclusion of particular rules. The approach reflects a desire to be useful for multiple projects and papers, with resulting measures being used as dependent variables (e.g., model how welfare policy is driven by the racial composition of the caseload), and as independent variables (e.g., how state policies impact how frequently clients are sanctioned). These seven steps can be applied to immigration, education, and other policies with high levels of complexity or accountability. However, the illustrations below all emerged in the context of a specific research agenda concerning the racialization of welfare policy in the United States.

Step 1: Identify All Choice Sets Imposed by, or Allowed by, Federal Authorization

The approach begins with the federal structure of TANF, which created a number of explicit *choice sets* for each state. Guided by prior research and the organizational structure of the Welfare Rules Database, Ojeda and colleagues (2017) identified 20 choice sets, listed in Table 1, that include more than 500 rules.[1]

Table 1. Major TANF Choices Sets Confronting State Governments

	Number of Rules
Does the state try to divert the family from applying for benefits?	18
What kinds of families are potentially eligible?	19
Which people within a family are potentially eligible?	82
Amount of assets a family can have and still be eligible?	8
How is income counted?	67
Amount of income a family can have and still be eligible?	71
If family passes all eligibility tests, what factors determine the amount of their benefit?	9
What are a recipient family's child support requirements?	26
What are a recipient family's behavioral requirements?	47
What are a recipient family's work-related activities requirements?	17
How soon must one satisfy work activity requirements in order to receive cash assistance?	1
Who is exempt from having to work?	15
How many hours must a recipient work each week?	2
How severe are the sanctions for failure to work?	31
How many months can a recipient stay on the welfare rolls?	7
Who is allowed to exceed the stated time limit and for how long?	34
Are there time limits pertaining to waiting periods before or between months of eligibility?	38
Who is allowed to exempt a month on TANF from counting against to the lifetime limit?	17
Are there additional requirements and rules for parents under the age of 18?	13
What happens after a family no longer receives benefits?	39
Total	561

Note: In many states, a different version of these rules apply to different kinds of recipients (e.g., different rules based on the age of the parent, presence of young children, etc.)

Step 2: Determine the Choice Sets Relevant to the Research Project

Few studies seek a comprehensive map of a policy regime as their primary goal. For example, Berkman, Honaker, Ojeda, and Plutzer (2013a) were interested in explaining differential rates of sanctioning, and therefore sought to identify *all* the relevant choice sets in the state welfare policies by asking three questions: (i) What rules would have to be violated in order for a case to be sanctioned or closed? (ii) What situations excuse TANF recipients from meeting some or all of these requirements? (iii) How severely are recipients penalized for a rule violation? They then identified all rules that helped answer any of these questions. For each rule identified, they then selected the entire choice set that included this rule.

In contrast, the current illustration focuses on the consequences of minority caseload, a central variable in Soss et al.'s Racial Classification Model (2011). This model draws upon certain well-established racial stereotypes and how these stereotypes lead policymakers to particular policy outcomes. Two papers (Whitesell, 2015, 2017) focused on highly sexualized stereotypes that envision African Americans as sexually promiscuous (Collins, 2002; Neubeck & Cazenave, 2001), "breeders" (Mink, 1998; Smith, 2007), and socially irresponsible (Sparks, 2003). These are especially relevant to choice sets concerning family formation (the family cap), and good parenting (cash assistance dependent on child immunization, regular school attendance).

Other stereotypes mark African Americans as work averse, lazy, and content to be economically dependent on the state (Collins, 2002; Gilens, 1999; Johnson, Duerst-Lahti, & Norton, 2007; Neubeck & Cazenave, 2001; Sparks, 2003). These have particular relevance for sets examined by Ojeda et al. (2017), and five of their choice sets will serve as exemplars for the remainder of this paper:

1. What activities count as work?
2. Who is exempt from having to work?
3. How many hours must a recipient work?
4. How severe are the sanctions for failing to work?
5. How many months can a recipient remain on the welfare rolls?

These directly address the notion that work requirements must be strict, that time on welfare is limited in order to motivate job search and retention, and that states should eliminate financial incentives for those who are lazy and content to remain dependent on public assistance.

Step 3: Identify Every Codable Rule in Every Selected Choice Set; Determine if the Choice Set Includes Subcategories

If a choice set is identified as relevant to the research project, include *every* rule in that set rule in the data collection. This eliminates any subjective judgment about which rules are most "important" or relevant and avoids inappropriate scale purification. Second, assess whether the choice set comprises multiple subscales. For example, consider the central choice set for a "workfare" policy—the seemingly simple policy decisions concerning which activities count as "work" in the context of a "workfare"-oriented policy.

The basic choice set, summarized in Table 2, includes 17 kinds of activities that states may designate as "work."[2] However, two major complications weigh against simply adding up the number of allowed activities. First, many states have one set of allowable activities for the most typical recipients, and a second, third, or even a fourth list of allowable activities for others. For example, many states provide more flexibility to recipients with young children than to those with school-age children. Thus, some states seemingly have 17 work activity rules, some 34, 51, or even 68; this complication is addressed in Step 6. Ignoring that for the moment, it is critical to recognize that only some of these activities count when a state seeks to demonstrate that a high percentage of its caseload is "working." Specifically, the last four activities in Table 2 can never count as meeting the federal government's definition of "work" as operationalized in the state's Work Participation Rate (WPR). This is important because if a state's WPR is too low, it will lose part of its federal block grant. These last four activities represent a degree of extra flexibility that states can extend to welfare recipients but at the cost of placing federal funds at risk if too

Table 2. Activities that States Can Designate to Count as "Work"

Federally recognized core activities	
	Job search is an allowable activity
	Community Work Experience or Alternative Work Experience (CWEP/ AWEP) are allowable activities
	Providing child care for others is an allowable activity
	Community service is an allowable activity
	Job readiness activities are allowable activities
	On-the-job training is an allowable activity
	Unsubsidized employment is an allowable activity
	Work supplementation or subsidized employment are allowable activities
Federally recognized noncore activities	
	High school attendance and or work toward a GED are allowable activities
	English as a Second Language classes are allowable activities
	Basic or remedial education are allowable activities
	Postsecondary education is an allowable activity
	Job skills training is an allowable activity
Activities not recognized by federal government as counting as "work"	
	Job development and job placement are allowable activities
	Self-employment is an allowable activity
	Counseling is an allowable activity
	Life skills training is an allowable activity

many recipients use these activities to maintain their eligibility. Given the importance of the WPR as an accountability measure, these four activities comprise their own, distinct, choice set.

Similarly, federal regulations allow the activities in the middle panel of Table 2 to count toward the WPR only after a recipient completes 20 or more hours of work that week in one of the eight "core" activities. In this light, these 17 rules actually constitute three sets of choices: how many of the eight "federal core" activities and how many "noncore" activities should be allowed, and whether to allow one or more of the four activities that the federal government does not recognize. The different connections between each choice set and TANF's financial accountability regime leads to a theoretical expectation that states would enact these choice sets differently. This illustrates how an apparent choice set might be composed of two or more sets that each have different policymaking antecedents and different impacts on citizens (not only welfare recipients, but also employers, educational institutions, and others).

Step 4: Determine If Each Coded Rule Is Independent of or Contingent on the Others; If So, Preprocess and Combine Rules when Necessary

In the example above, the flexibility of a state's work requirements is a positive function of the number of "work" options afforded to program participants, and if a state allows one activity, this does not preclude the acceptance of another. However, whenever the presence of a code of "1" for a particular rule is contingent on the code of another rule, they are not statistically independent, thereby violating assumptions implicit in most commonly employed data-reduction techniques. Thus, we recommend that researchers decide how to deal with such cases *before* constructing scales.

For example, consider the severity of sanctions, a central focus of nearly every empirical analysis of TANF. The particulars of the 30 different rules governing sanctions complicate the construction of a summary measure. In most states, sanctions escalate such that "strike one" results in a warning or a mild sanction, and a penalty for "strike two" is more severe, with escalating penalties until a recipient is completely cut off or even banned for life. Table 3 illustrates this complexity by showing the progression of sanctions across four different states in a typical year.

Necessarily, if the initial penalty is very severe, the values of the later penalties are constrained, which can shrink or reverse correlations between the "strike one" and later rules. These effects will have pernicious consequences for scale construction if not recognized and accommodated.

The absence of statistical independence is a common challenge in measurement more generally. Standardized tests, for example, often present a short reading passage and ask students to answer 3–5 questions about the passage; these questions comprise a testlet. Educational psychologists have long recognized that the resulting data will violate the assumption of independent errors because a feature of the reading assignment can impact student answers to all the questions that follow. While it is possible to model these errors with a random effect for each testlet (Li, Bolt, & Fu, 2006), it is more common to combine answers of the individual items into a testlet score, which will be independent of answers to other questions in the test (Wainer, 1995).

Likewise, we suggest that nonindependent codes, such as the progression of sanctions, be preprocessed. Usually, researchers can combine two or more complementary, or logically contingent, rules into meaningful numeric or ordinal scales. Ojeda and colleagues (2017), for example, created an ordinal code for each state's initial sanction so that the four states above would receive initial severity scores of 0, ½, and 1, reflecting that TANF recipients who violate their work requirements receive only a warning in South Dakota, a partial reduction in cash benefits in New Hampshire and Nevada, and an immediate 100 percent reduction in Wyoming. But the initial sanction only tells part of the story. TANF recipients in New Hampshire

Table 3. Sanction Severity in Four Illustrative States, 2000

	The Reduction in Benefits Is:				
	1st Violation	*2nd Violation*	*3rd Violation*	*4th Violation*	*5th Violation*
New Hampshire	Partial	Partial	Partial	Partial	Partial
South Dakota	Zero (warning)	Partial	Partial	–	–
Nevada	Partial	Partial	Total	–	–
Wyoming	Total	–	–	–	–

continue to receive penalties that reduce their benefits after multiple violations, while Nevada participants escalate to a 100 percent reduction on the third offense. Since no state issues a warning as its final penalty, they created a final sanction code of zero for a partial sanction and one for a 100 percent penalty—so these states receive codes of 0, 0, 1, 1. When these codes are combined, South Dakota is most generous (0 + 0 = 0), followed by New Hampshire (½ + 0 = ½), Nevada (½ + 1 = 1½), and Wyoming (1 + 1 = 2).

In combining information from five codable rules into two intermediate measures, this procedure eliminates some missing data challenges (Wyoming is missing for four of the five rules), reduces statistical dependency, and arrives at an ordinal measure that seems to validly rank the states from least to most severe. And by counting Wyoming's severe sanction twice, as both the initial and "worst" sanction, this approach accounts for the absence of a gradual escalation.

An analogous procedure was applied to three additional sanction rules: whether the state provides a conciliation process whereby a recipient can correct their noncompliant status before the sanction is imposed, whether the sanctions escalate to case closure, and whether case closure is accompanied by a lifetime ban as a final sanction.[3] These were then combined and averaged to create a summary measure of sanction severity on a zero–one scale (see Appendix for details in Supporting Information).

Step 5: Use Theory and Substantive Knowledge to Specify the Dimensions that Should Define Each Choice Set and Each Item

Steps 1–4 lead scholars to identify the relevant choices sets, select the choice sets relevant to the study, enumerate every rule in each choice set, and preprocess rules that are statistically or logically dependent on others. In step 5, researchers must assign to each rule and choice set polar adjectives such as flexible/inflexible, forgiving/punitive, or lenient/strict and do so systematically and transparently. This may appear to be trivial, but many works display confusion in this regard. For example, Fellowes and Rowe (2004) create an index intended to measure the flexibility of a state's welfare requirements, but this index includes several measures of sanction severity. A mismatch of polar adjectives is a good indication of face invalidity.

Determining the polar direction of a specific rule can also be in error. In determining direction, scholars typically use one of several common approaches. One used in both psychometrics and legislative roll call analysis is to select one key rule and designate its polar valence (e.g., liberal, generous, racist) and then allow an algorithm to determine the valence of all other elements in relationship to the designated anchor case, thereby uncovering latent dimensions. This can be helpful when the researcher is agnostic to the inclusion of items. For example, many policy descriptions have a category for "other" and it may not be clear whether the existence of this category makes the policy more or less generous, more or less flexible, and so on. In the preceding example, the sanction severity scale includes indicators for the presence of a conciliation (appeals) process and an indicator for the severity of the initial sanction. Most data-reduction algorithms will incorrectly flag these as "reverse coded" and we re-emphasize that allowing a measurement model to guess at the direction is a poor choice whenever the scale includes complementary rules or when legislative trade-offs are likely.

A second approach is to assess the valence of each item from the perspective of those subject to the rule or some other theoretically informed criterion. For example, Step 4 described the negative correlation between a state imposing case closure and a state imposing a partial benefit reduction. Both are clearly negative outcomes for a recipient and intended as incentives to get and retain work. Based on this knowledge, these should both be coded and weighted in the same direction *even if a data reduction model suggests opposite signs*. This item-by-item judgment also allows for some policy rules within a choice set to have a different expected direction relative to the others (analogous to a reverse-worded survey question).

Step 6: Determine How to Weight and Combine Elements into a Summary Scale

Once choice sets are defined, nonindependent items identified and combined, and the direction of each policy element is determined, it is time to combine the items into a summary score for the choice set. The simplest case is that of unit weighting—count each element in a choice set the same. There are several approaches to calculate weights for a weighted sum, but in many instances, researchers will lack the information and theory to implement these correctly.

Option 1: Weight Based on Data-Reduction Models. One way to differentially weight policy elements is via scaling methods that produce estimates of the value of a latent variable. IRT, factor analysis, and similar methods give greater weight to rules that display the greatest centrality (as indicated, for example, by the mean inter-item correlation or covariance). Applying these methods to state ($N = 51$) or state-year data is rarely statistically sound, however, because the number of estimated parameters quickly approaches the sample size. The minimum sample size for IRT models is given by (*Rules**[*Rules* + 1])/2 (Wirth & Edwards, 2007) which means that an IRT model with 10 rules requires 55 states and those with 17 rules (as in the case of work activity requirements) would require $n \geq 136$. Adding cases by utilizing state-years seemingly solves the degrees of freedom problem, but only if one assumes temporal independence. If temporal dependence (a state's law in one year is likely to be the same as in the prior year) is ignored, the effective degrees of freedom will be far fewer than the apparent number of state-years, and this will result in unstable estimates that vary widely in magnitude. For example, a conference paper reporting preliminary analysis of these rules (Berkman et al., 2013b) report state-level IRT models that imply that a partial reduction in cash benefits should be weighted more than 10 times as much as a total reduction in benefits, a conclusion that fails the commonsense test. One indication of this problem is when item discrimination parameters approach the theoretical maximum, for example when they exceed a value of four on a *Z*-scale. Our recommendation is to use caution when using such models as we think that the model assumptions will rarely be met in data sets used for comparative subnational or cross-national data sets unless the number of rules is very small.

It is sometimes possible to adapt measurement models to the needs of policy measurement. Realities such as temporal dependence (e.g., Walls & Schafer, 2006) and complementary items (Li et al., 2006) can be modeled but these fixes have been developed using larger samples, which can accommodate the many additional parameters specified. Alternatives that generate weights simultaneously with the estimation of effects between the scale and another (independent or dependent) variable remain controversial (Bollen & Diamantopoulos, 2017). Thus, the challenges of adapting these methods to applications such as state policy measurement is currently beyond the state of the art. Until policy scholars develop methods that more efficiently use the data, approaches like the one detailed here will represent the best we can do.

Option 2: Weight by the Number of Individuals Impacted. When welfare rules vary for different segments of the welfare caseload, it may be prudent to weight items in proportion to the number of families impacted by each rule variation—especially if the measure is used to predict outcomes among the recipient population. This principle is applied, for example, by Monogan's (2013, 2018) measurement of state immigrant policy. He first created an ordinal measure of *scope* running from "1" (law is merely symbolic) to "4" (the law directly affects "immigrants' ability to reside in a state"). Second, coders classified each law as welcoming or hostile to immigrants (tone). Monogan (2013) aggregates by taking a *weighted* ratio of welcoming laws to hostile laws, with greater weight assigned to laws of greater scope (these are further disaggregated by Filindra, 2018, and Monogan, 2018, into hostile and welcoming laws analyzed in separate models). That is, tone was weighted by the scope of the law, before summing up to get a state-year score.

This approach can simplify the multiplication of work activity rules. As noted earlier, states typically identify a number of core, noncore, and nonfederal activities that recipients can use to satisfy their requirement to "work." When states have different rules for different classes of recipients, they might permit four of the core activities for high school graduates who are parents of school-age children, six activities for high school graduates who have small children, and all eight activities for high school drop-outs with young children. In terms of scoring the state in terms of its work flexibility, we can take the simple mean of these rules ([4 + 6+8]/3 = 6) or, analogous to Monogan, weight these in proportion to the state's caseload that is governed by each rule. We calculated it in both ways and the results are essentially the same. But this will not typically be the case.

Option 3: Unit Weighting. The simplest approach is to simply "add 'em up" by counting each element equally (equivalent to assigning all included rules a weight of "1" and any excluded rules a weight of "0"). Unit weighting has a long tradition in state politics research (e.g., Squire's index of legislative professionalism), partly because "add 'em up" scales are usually highly correlated with measures that assign magnitudes to the weights (e.g., see Bowen & Greene, 2014; and nearly the same as calculating the geometric mean, see Bjerre et al., 2018). Unit weighting does require some subjective decision making by the investigator, however. For example, most scholars rescale all items to an arbitrary scale having a fixed minimum and maximum scale such as 0–100 or 0–1. Others may transform to a *Z*-distribution (mean of 0, and SD of 1), or percentile rank before calculating a summary score.

Allowing each element to contribute the same amount may seem suboptimal. However, an *a priori* decision to use unit weights avoids capitalizing on chance from measurement models calculated from small samples. Weights derived from small-sample analyses are not likely to be optimal for other small samples and may thereby hinder replication and cumulation of findings. Especially with small samples, there is no perfect solution to the choice between a simple additive scale and one that gives greater weight to some policy elements and less to others. Unit weights were employed to create the seven summary measures described in Table 4.

Step 7: Assess the Validity of Policy Measures

Every empirical analysis rests on the validity of its variables. Step 5 addresses *content validity* explicitly as it requires researchers to assign polar adjectives to binary codes, ordinal codes, and entire choice sets. Step 7 extends this to illustrate how to provide evidence of discriminant validity and criterion validity. This is illustrated by focusing on the five choice sets initially identified—now seven measures, due to the disaggregation of allowable work activities based on how they count in the federal accountability policy.

Discriminant Validity. Measures intended to reflect different theoretical concepts should be statistically distinct. In this light, we examine the pairwise correlations among the seven measures. Table 5 shows that these various policies are quite dissimilar, with the highest correlation being 0.35. The correlation matrix also provides evidence of policy trade-offs. The small, negative, correlation ($r = -0.25$) between more generous time limits (bottom row) and more generous sanctions (column 4) contradicts the idea that all elements of a state's welfare policy reflect a single political ideology.

Criterion Validity. Criterion validity refers to the ability of a measure to predict outcomes in accord with theoretical expectations. In the substantive paper that motivated the measurement of these choice sets, Ojeda and colleagues (2017) argued that if policymakers held negative stereotypes of African Americans, then as the Black share of the caseload increased, the severity of sanctions would also increase.

However, if policymakers hold stereotypes that African Americans are lazy and either unwilling or unable to maintain employment, then a large Black caseload would make it difficult to achieve federally mandated targets for the percentage of recipients who are working, placing the state's block grant in jeopardy. This would lead policymakers to define "work" in the most flexible way possible, but only those work activities included in accountability formulas.

These observations lead to a specific prediction: As the Black share of a state's welfare caseload rises, a state should *increase* the number of federal core activities

Table 4. Descriptive Statistics for Measures of Seven Choice Sets

		Summary Statistics 1997–2006			
Abbreviation	Description	Min	Max	Mean	Std Dev
Activities allowed—core*	Number of federal core activities recipients may use to satisfy work requirements	2	8	5.99	1.44
Activities allowed— noncore*	Number of federal noncore activities recipients may use to satisfy work requirements	0	5	4.34	0.94
Activities allowed —nonfederal*	Number of nonfederal activities recipients may use to satisfy work requirements	0	4	1.67	1.07
Activities exemptions*	Number of recipient types who are exempt from having to work	1	8	4.81	2.15
Hours of work required*	Ordinal measure with 1 = under 20 hours, 2 = exactly 20 hours, 3 = 21–29 hours, 4 = exactly 30 hours, 5 = more than 30 hours	1	5	3.31	0.86
Sanction severity	Mean of three initial and four final sanction severity indicators	0	1	0.44	0.24
Time limits	–1 = stricter than the federal maximum, 0 = exactly 60 months, +1 = more generous than federal maximum	–1	1	0.01	0.56

*These measures are averages of the requirements for different types of clients.

Table 5. Inter-Correlations (Pearson *r*) Among Choice Set Measures (1997–2006)

	Core	Noncore	Nonfederal	Activities Exemptions	Hours of Work Required	Sanction Severity	Time Limits
Activities allowed—core	1.00	–	–	–	–	–	–
Activities allowed—noncore	0.11	1.00	–	–	–	–	–
Activities allowed—nonfederal	0.13	0.35	1.00	–	–	–	–
Activities exemptions	0.10	0.11	−0.14	1.00	–	–	–
Hours of work required	0.03	0.21	0.12	0.12	1.00	–	–
Sanction severity	−0.20	0.14	0.01	0.05	−0.13	1.00	–
Time limits	0.04	0.00	−0.02	−0.25	0.12	−0.11	1.00

that qualify for "work." Thus, race should be related to a more flexible and generous policy, but only for core activities. This is born out in the data. Figure 2 reports the key results of models controlling for year fixed effects, state random effects, and a range of demographic and political controls (based on analyses reported in Ojeda et al., 2017). All policy subsets are rescaled so that the theoretical minimum is zero and theoretical maximum is one, to aid interpretation. The coefficient estimates show that as the black share of the caseload rises, sanctions get *tougher*, but it is *easier* to satisfy the federal core work activity requirement. Race has no effect on choice sets that are less relevant or irrelevant to federal accountability rules.

These results show that the decision to break up allowable work activities into three scales based on how they figure in federal accountability calculations produced more valid measures than would have been achieved with a single, 17-point scale. The low correlations among our constructs suggests that they pass the test of discriminant validity and the predicted, but opposite, correlations with the racial composition of a state's caseload is a strong indicator of criterion validity.

Comparisons with Other Measurement Strategies

The steps we recommend are rooted in principles that should lead to measures that are comprehensive, replicable, and whose effects are interpretable. We provide evidence for this by comparing our measures to the three most influential alternatives.

The first of these are the flexibility and eligibility indexes reported in Fellowes and Rowe's widely cited paper (2004). They take 40 rules and combine them into an eligibility index and a flexibility index. Other authors (e.g., Kim & Fording, 2010; Reingold & Smith, 2012) have also used these measures, as independent or control

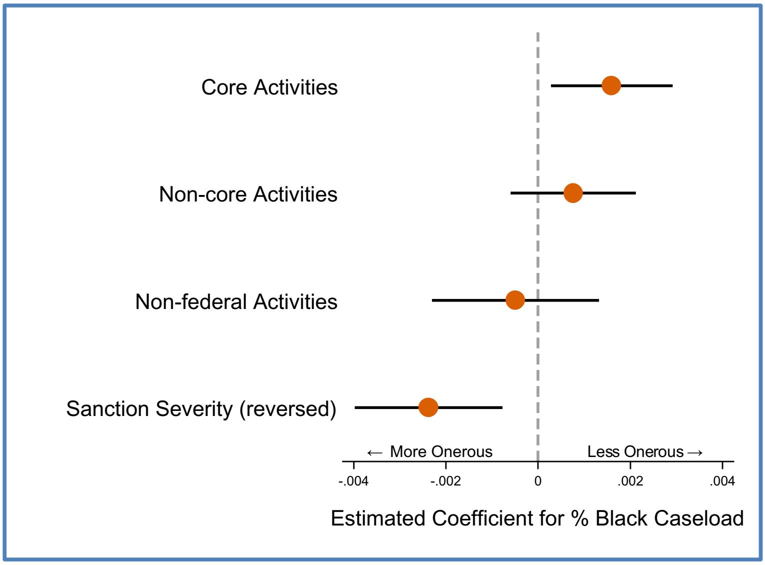

Figure 2. Effect of Black TANF Caseload Percentage on Four Welfare Rule Choice Sets (1997–2006).

variables. Following the documentation in their 2004 article, we replicated the measure for the period before the Deficit Reduction Act (1997–2006).

Their flexibility index comes closest to our focus on work requirements and we will focus on that. It is immediately obvious that the scale construction involved considerable choice among all the possible rules that might reflect the concept of "flexibility": their index includes six of eight rules that identify types of individuals who are exempt from working, one rule pertaining to work hours, three pertaining to sanctions, and just two of the seventeen allowable work activities. It thus spans four of the choice sets we identified.

The Fellowes and Rowe (2004) flexibility index is modestly correlated with our measures of sanction severity (−0.37), time limits (r = 0.28), and noncore work activities (r = 0.28). Most critically, it is uncorrelated (r = −0.09) with the state's flexibility in terms of the number of core activities—those which must be satisfied for at least 20 hours each week. Thus, their index does not in any way reflect the flexibility of work activities from the perspective of the client. The unexplained omission of all eight core requirements has the result of eliminating rules that are positively correlated with the racial composition of the caseload, leading Fellowes and Rowe to incorrectly conclude that the black share of the caseload is negatively correlated with flexibility.

The second most prominent examination of the effect of state racial caseload on welfare rules is that of Soss et al. (2001). They examined four "get tough" policies. Like Ojeda and his colleagues (2017), they find a strong relationship between Black caseload and sanction severity, but they find no effect at all of race on what they call "stricter work requirements." As noted above, "strict" can apply to the number of hours one must work, how easy it is to secure an exemption, or whether many activities (education, training, community service) can count as "work." Like Fellowes and Rowe, Soss et al. (2011) include none of these rules and therefore fail to identify the consistent link between race and more flexible work requirements.

The third most prominent effort is that of De Jong and colleagues (2006). Like Fellowes and Rowe (2004), they begin with only a tiny fraction of all possible rules. They describe the process as first enumerating 78 rules that had meaningful cross-state variation, then recoding them so that all are scored with zero as most lenient and one (or higher, in the case of ordinal measures) as most stringent. After that step, they followed purification principles of Classical Test Theory that apply to reflective latent variable models: "we used bivariate correlation analysis … to explore which items together might represent underlying policy dimensions." Rules that were not eyeballed as being intercorrelated were dropped. The retained rules were subject to a second round of item purification which only retained rules with a factor loading of 0.4 or greater; these two steps eliminated 35 out of 78 rules.

The process was not transparent, so it is impossible to discern which rules failed the initial interstate variation test and which were eliminated in each purification step. Their archived summary scores, however, provide some hints. Their "Activities Requirements" dimension is correlated at a level of 0.08 with Soss et al.'s (2011) "strictness" measure, 0.21 with Fellowes and Rowe's (2004) flexibility index, and correlates at a level of 0.45 with our federal core activities requirements (but uncorrelated with the noncore and nonfederal scales). As a result, the De Jong et al. (2006) measure also shows the same correlation with race reported above, but there is nothing in the construction of the scale that would provide a hint that this was due to fiscal and federal accountability concerns.

Summary and Discussion

Like education, environmental regulation, economic development, or other complex policy regimes, welfare policy and immigration policy are multifaceted. Scholars must ensure that preconceptions do not influence research designs in ways that give rise to self-fulfilling confirmations. Measuring welfare or immigration policy based primarily on features that have received the most attention runs the risk of selecting policy elements already suspected of having disparate effects due to factors like conscious or unconscious racism, nativism, or religious prejudice. Rather, researchers should recognize that racialization takes many forms. In the example explored here, the less salient details of fiscal accountability and race interacted in theoretically predictable ways. But fiscal accountability has received less attention than the more ideologically pitched rhetoric of the welfare queen. To avoid these blind spots, we recommend that scholars cast the widest possible net, identify all possible choice sets, and then seek to include every single rule within a choice set in the measurement process.

This advice applies equally to the other core theme of this special issue: immigration policy. Often, policies that apply to immigrants and other noncitizens in the United States are embedded in larger choice sets. With TANF, for example, a number of eligibility rules—those in our choice set, "Which people within a family are potentially eligible?" apply different criteria to citizens and noncitizens. But we do not recommend examining these particular rules in isolation from the closely related policy choices states make on eligibility more generally. Likewise, laws that treat immigrants differently from citizens in terms of eligibility for social insurance (e.g., workman's compensation), health care, professional licensing, eligibility for in-state college tuition, and many others (Morse, Mendoza, & Mayorga, 2016) should be viewed in the broader context of the choice sets germane to those policies; this is because enacted immigrant and immigration policies are also impacted by trade-offs, compromises, and fiscal concerns and these can only be discerned in the context of all the relevant choice sets confronting policymakers.

We also have warned about some of the dangers of trying to reduce a large number of policy codes to a manageable number using data-reduction methods whose main assumptions derive from psychological measurement. Measuring personality traits or mental illness presumes stable underlying conditions. In contrast, policy is produced by people in institutions who have many, often contradictory, motivations. We contend that the underlying policy generation process will not typically correspond to the assumptions of measurement models and the burden of demonstrating this lies with the researcher. Identifying the key dimensions and labeling policies' directions on these dimensions (Step 5) is a critical step to ensure that data-reduction methods score policies in the correct direction. In the domain of immigration policy, this principle is nicely illustrated by Monogan's (2013) coding of policies as welcoming or hostile before aggregation.

Our empirical focus, the "workfare" aspects of TANF, revealed clear examples of statistical dependence, and of policy trade-offs that can blunt the impact of ideology or stereotypes. To account for these, we proposed a series of specific measures that reflected the specific choices confronting policymakers and urge scholars to reduce the opportunities to make subjective judgments and to provide sufficient transparency so that all such decisions are replicable. This process resulted in an important qualification to the idea that increased Black caseloads always make policy less generous, less flexible, and more onerous. And it contributed to theoretical development inasmuch as our work provided an important addition to the Racial Classification Model. These gains were only possible because we considered a wider range of policy dimensions (choice sets), considered every rule in each choice set, and because we did not automatically exclude policy rules that "do not scale well" according to commonly used criteria. We think such nuanced findings are likely to emerge in other research programs as well.

Eric Plutzer is professor of political science and sociology at Penn State University.
Michael B. Berkman is the director of the McCourtney Institute for Democracy, and professor of political science at Penn State University.
James Honaker is research associate at the Center for Research on Computation and Society and at Harvard University.
Christopher Ojeda is an assistant professor of political science at the University of Tennessee.

Ann Whitesell is an assistant professor of political science at Ohio Northern University.

Notes

1. Of course, identifying the choice sets of complex policy regimes will not always be so straightforward. Policies that are less politically salient may not garner attention from nonprofit groups, such as the Urban Institute, that are collecting data on policy variation across the states. Here we recommend that researchers draw on the law itself, prior research, experts in the field, and the relevant bureaucrats to determine the scope and depth of the law.
2. The Deficit Reduction Act of 2005 altered the classification of three rules, effective in 2007. For the purposes of illustration, we are using the policy formulas in place prior to this change.
3. Extensive missing data prevented us from utilizing an additional sanction severity measure concerning whether other benefits (e.g., food stamps, Medicaid) are also reduced due to this initial or subsequent violation.

References

Avery, James M., and Mark Peffley. 2005. "Voter Registration Requirements, Voter Turnout, and Welfare Eligibility Policy: Class Bias Matters." *State Politics & Policy Quarterly* 5 (1): 47–67.

Berkman, Michae, James Honaker, Christopher Ojeda, and Eric Plutzer. 2013a. *Written Policy and Client Outcomes in Temporary Aid to Needy Families: The Impacts of Rules, Race, Institutions and Context*. Paper presented at the APSA Annual Meeting, Chicago, August 27–30, 2013.

———. 2013b. *Measuring State Welfare Policies*. Presented at the State Politics and Policy Annual Meeting, Iowa City, Iowa.

Bjerre, Liv, Friederike Römer, and Malisa Zobel. 2018. "The Sensitivity of Country Ranks to Index Construction and Aggregation Choice: The Case of Immigration Policy." *Policy Studies Journal*. https://doi.org/10.1111/psj.12304

Bollen, Kenneth A., and Adamantios Diamantopoulos. 2017. "In Defense of Causal-Formative Indicators: A Minority Report." *Psychological Methods* 22 (3): 581–96.

Bollen, Kenneth, and Richard Lennox. 1991. "Conventional Wisdom on Measurement: A Structural Equation Perspective." *Psychological Bulletin* 110: 305–14.

Bowen, Daniel C., and Zachary Greene. 2014. "Should We Measure Professionalism with an Index? A Note on Theory and Practice in State Legislative Professionalism Research." *State Politics & Policy Quarterly* 14: 277–96.

Cnudde, Charles F., and Ronald J. McGrane. 1968. "Party Competition and Welfare Policies in the American States." *American Political Science Review* 62: 1220–31.

Collins, Patricia Hill. 2002. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. New York: Routledge.

Dawson, Richard E., and James A. Robinson. 1963. "Inter-Party Competition, Economic Variables, and Welfare Policies in the American States." *Journal of Politics* 25: 265–89.

De Jong, Gordon F., Deborah Graefe, Shelley K. Irving, Tanja St. Pierre. 2006. "Measuring State TANF Policy Variations and Change After Reform." *Social Science Quarterly* 87 (4): 755–81.

Diamantopoulos, Adamantios, Petra Riefler, and Katharina P. Roth. 2008. "Advancing Formative Measurement Models." *Journal of Business Research* 61: 1203–218.

Dye, Thomas R. 1984. "Party and Policy in the States." *The Journal of Politics* 46: 1097–116.

Fellowes, Matthew C., and Gretchen Rowe. 2004. "Politics and the New American Welfare States." *American Journal of Political Science* 48 (2): 362–73.

Filindra, Alexandra. 2013. "Immigrant Social Policy in the American States: Race Politics and State TANF and Medicaid Eligibility Rules for Legal Permanent Residents." *State Politics & Policy Quarterly* 13 (1): 26–48.

———. 2018. "Is 'Threat' in the Eye of the Researcher? Theory and Measurement in the Study of State-Level Immigration Policymaking." *Policy Studies Journal*. https://doi.org/10.1111/psj.12264

Fording, Richard C. 1997. "The Conditional Effect of Violence as a Political Tactic: Mass Insurgency, Electoral Context and Welfare Generosity in the American States." *American Journal of Political Science* 41: 1–29.

Fording, Richard C., and William D. Berry. 2007. "The Historical Impact of Welfare Programs on Poverty in the American States." *Policy Studies Journal* 35 (1): 37–60.

Gilens, Martin. 1999. *Why Americans Hate Welfare: Race, Media, and the Politics of Antipoverty Policy*, Chicago: University of Chicago Press.

Goodman, Sara Wallace. 2018. "Indexing Immigration and Integration Policy: Lessons from Europe." *Policy Studies Journal.* https://doi.org/10.1111/psj.12283

Graefe, Deborah R., Gordon F. De Jong, and Shelley K. Irving. 2006. "The Whole is the Sum of its Parts: Theory, Technique, and Measurement Science Applied to TANF Rules." *Social Science Quarterly* 87 (4): 818–27.

Hill, Kim Quaile, and Jane E. Leighley. 1992. "The Policy Consequences of Class Bias in the State Electorates." *American Journal of Political Science* 36 (2): 351–65.

Hill, Kim Quaile, Jan E. Leighley, and Angela Hinton-Andersson. 1995. "Lower-Class Mobilization and Policy Linkage in the U.S. States." *American Journal of Political Science* 39 (1): 75–86.

Hofferbert, Richard I. 1966. "The Relation Between Public Policy and Some Structural and Environmental Variables in the American States." *American Political Science Review* 60 (1): 73–82.

Jennings, Edward T. 1979. "Competition, Constituencies, and Welfare Policies in American States." *American Political Science Review* 73 (2): 414–29.

Johnson, Cathy Marie, Georgia Duerst-Lahti, and Noelle H. Norton. 2007. *Creating Gender: The Sexual Politics of Welfare Policy*. Boulder, CO: Lynne Rienner Publishers.

Key Jr., V. O. 1949. *Southern Politics in State and Nation*. New York: Knopf.

Kim, Byungkyu, and Richard C. Fording. 2010. "Second-Order Devolution and the Implementation of TANF in the US States." *State Politics & Policy Quarterly* 10 (4): 341–67.

Li, Yanmei, Daniel M. Bolt, and Jianbin Fu. 2006. "A Comparison of Alternative Models for Testlets." *Applied Psychological Measurement* 30 (1): 3–21.

Lieberman, Robert C. 2001. *Shifting the Color Line: Race and the American Welfare State*. New York: John Wiley & Sons.

Lockard, Duane. 1959. *New England State Politics*. Princeton, NJ: Princeton University Press.

Manza, Jeffrey. 2000. "Race and the Underdevelopment of the American Welfare State." *Theory and Society* 29 (6): 819–32.

McKernan, Signe-Mary, Jen Bernstein, and Lynne Fender. 2005. "Taming the Beast: Categorizing State Welfare Policies: A Typology of Welfare Policies Affecting Recipient Job Entry." *Journal of Policy Analysis and Management* 24 (2): 443–60.

Mettler, Suzanne, and Joe Soss. 2004. "The Consequences of Public Policy for Democratic Citizenship: Bridging Policy Studies and Mass Politics." *Perspectives on Politics* 2 (1): 55–73.

Mink, Gwendolyn. 1998. *Welfare's End*. Ithaca, NY: Cornell University Press.

Monnat, Shannon M. 2010. "The Color of Welfare Sanctioning: Exploring the Individual and Contextual Roles of Race on TANF Case Closures and Benefit Reductions." *The Sociological Quarterly* 51 (4): 678–707.

Monogan, James E. 2013. "The Politics of Immigrant Policy in the 50 US States, 2005–2011." *Journal of Public Policy* 33 (1): 35–64.

Monogan, James. 2018. "Studying Immigrant Policy One Law at a Time." *Policy Studies Journal*.

Morse, Ann, Gilberto Soria Mendoza, and Jennifer Mayorga. 2016. *Report on 2015 State Immigration Laws*. Washington, DC: National Conference of State Legislatures.

Neubeck, Kenneth J., and Noel A. Cazenave. 2001. *Welfare Racism: Playing the Race Card against America's Poor.* New York: Routledge.

Ojeda, Christopher, Anne Whitesell, Michael Berkman, and Eric Plutzer. 2017. *Federalism and the Racialization of Welfare Policy.* State Politics & Policy Conference, St. Louis, MO, June 1–3.

Orr, Larry L. 1976. "Income Transfers as a Public Good: An Application to AFDC." *American Economic Review* 66 (3): 359–71.

Pacheco, Julianna. 2013. "The Thermostatic Model of Responsiveness in the American States." *State Politics & Policy Quarterly* 13 (3): 306–32.

Piven, Francis Fox, and Richard Cloward. 1993. *Regulating the Poor.* New York: Vintage.

Plotnick, Robert D., and Richard F. Winters. 1985. "A Politico-Economic Theory of Income Redistribution." *The American Political Science Review* 79 (2): 458–73.

———. 1990. "Party, Political Liberalism, and Redistribution: An Application to the American States." *American Politics Quarterly* 18 (4): 430–58.

Reich, Gary. 2018. "One Model Does Not Fit All: The Varied Politics of State Immigrant Policies, 2005—16. " *Policy Studies Journal.* https://doi.org/10.1111/psj.12293

Reingold, Beth, and Adrienne R. Smith. 2012. "Welfare Policymaking and Intersections of Race, Ethnicity, and Gender in U.S. State Legislatures." *American Journal of Political Science* 56 (1): 131–47.

Schneider, Anne L., and Helen M. Ingram. 1997. *Policy Design for Democracy*. Lawrence: University Press of Kansas.

Schram, Sanford F. 2005. "Contextualizing Racial Disparities in American Welfare Reform: Toward a New Poverty Research." *Perspectives on Politics* 3 (2): 253–68.

Smith, Anna Marie. 2007. *Welfare Reform and Sexual Regulation*. New York: Cambridge University Press.

Soss, Joe, Richard C. Fording, and Sanford F. Schram. 2011. *Disciplining the Poor.* Chicago, IL: University of Chicago Press.

Soss, Joe, Sanford F. Schram, Thomas P. Vartanian, and Erin O'Brien. 2001. "Setting the Terms of Relief: Explaining State Policy Choices in the Devolution Revolution." *American Journal of Political Science* 45 (2): 378–95.

Sparks, Holloway. 2003. "Queens, Teens, and Model Mothers: Race, Gender, and the Discourse of Welfare Reform." In *Race and the Politics of Welfare Reform*, eds. Sanford F. Schram, Joe Soss, and Richard C. Fording. Ann Arbor: University of Michigan Press, 171–95.

Tweedie, Jack. 1994. "Resources Rather than Needs: A State-Centered Model of Welfare Policymaking." *American Journal of Political Science* 38 (3): 651–72.

Urban Institute. 2017. *Welfare Rules Database.* Washington, DC: The Urban Institute. http://wrd.urban.org/wrd/Query/query.cfm. Accessed September 29, 2017.

Wainer, Howard. 1995. "Precision and Differential Item Functioning on a Testlet-Based Test: The 1991 Law School Admissions Test as an Example." *Applied Measurement in Education* 8 (2): 157–86.

Walls, Theodore A., and Joseph L. Schafer, eds. 2006. *Models for Intensive Longitudinal Data*. Oxford: Oxford University Press.

Whitesell, Anne. 2015. *The Effects of Female and Minority State Legislator Incorporation on Neoliberal Paternalist Welfare Policy*. Paper presented at the 2015 MPSA Annual Meeting, Chicago IL.

———. 2017. *Who Represents the "Other"? The Influence of Organized Interests in State Welfare Policy*. Unpublished doctoral dissertation, The Pennsylvania State University.

Wirth, R. J., and Michael C. Edwards. 2007. "Item Factor Analysis: Current Approaches and Future Directions." *Psychological Methods* 12 (1): 58–79.

Wright, Gerald C. 1977. "Racism and Welfare Policy in America." *Social Science Quarterly* 57 (4): 718–30.

Supporting Information

Additional supporting information may be found online in the supporting information tab for this article.