STANFORD CENTER
on Poverty and Inequality Inequality.com

Measuring Social Class with Changing Occupational Classifications

Reliability, Competing Measurement Strategies, and the 1970-1980 U.S. Classification Divide

Pablo A. Mitnik
Center on Poverty and Inequality, Stanford University

Erin Cumberworth
Department of Sociology, Stanford University

November, 2016

The Stanford Center on Poverty and Inequality is a program of the Institute for Research in the Social Sciences (IRiSS). Support from the Elfenworks Foundation gratefully acknowledged.

Acknowledgment

The authors are grateful to Emily Beller and Michael Hout for sharing with them several algorithms to map occupations and other variables into Erikson–Goldthorpe–Portocarero classes.

Abstract

Periodic changes in occupational classifications make it difficult to obtain consistent measures of social class over time, potentially jeopardizing research on class-based trends. The severity of this problem depends, in part, on the measurement strategies used to address those changes. The authors propose that when a sample has been coded partly with one occupational classification and partly with another, Krippendorff's index α be used to identify the best strategy for measuring class consistently across the two classifications and to assess the reliability of the class measure employed in the final analyses. This index can be computed regardless of the metric of the class variable; it can be used to compare measures based on different class schemes or that use different metrics; and statistical inference is straightforward, even with a complex sampling design. The authors put the index to work in conducting a case study of the effects of the switch from the 1970 to the 1980 U.S. Census Bureau Classification of Occupations on the reliability of Erikson–Goldthorpe–Portocarero class measures. Their findings indicate that measurement strategies that seem a priori equally reasonable vary substantially in terms of their reliability, and that the bulk of this variation is accounted for by the extent to which the strategies rely on subjective judgments about the relationships between occupational and class classifications. Most importantly, as long as the best-performing measurement strategies are used, the switch in occupational classifications appears to be substantially less consequential than has been previously argued. A computer program made available as a companion to the paper makes estimation of Krippendorff's α, and statistical inference, very simple endeavors for nominal class variables.

Introduction

The study of long-term trends related to social classes—trends regarding a country's class structure, differences in political attitudes across classes, and intergenerational social mobility, to mention three prominent examples—requires that social class be measured consistently over time.[1] But consistent measures of social class can be hard to obtain. Class measures are typically derived from occupation variables (and sometimes other variables as well, such as self-employment status), and the classifications used to code occupations change periodically as they are improved conceptually and updated to account for changes in the occupational structure. When a dataset is coded partly with one occupational classification and partly with another, it can greatly affect social scientists' ability to study class trends. The severity of the problem depends, in part, on the nature and scope of the changes in the occupational classifications, but to a very large degree it also depends on the strategies used to deal with these changes.

There are usually multiple measurement strategies that can be used to "bridge" two occupational classifications, and a priori they often seem equally sensible. How do we know whether any given strategy produces a measure of social class reliable enough to study trends? How should we select among the available strategies? For nominal class measures, how can we determine the number of classes (or "granularity") to use in our analyses, given the potential trade-off between granularity and reliability? Similarly, if a class scheme favored on theoretical grounds is suspected to be less reliable than others

that are also acceptable, how can we establish the reliability cost of using the theoretically-preferred scheme?

Perhaps surprisingly, methodological tools for addressing these issues are not available, as social scientists have not developed any systematic approach for assessing the reliability performance of different measurement strategies when studying class-based trends across a change in occupational classifications. The main goal of this paper is to advance such an approach. We propose that a reliability index extensively used in the field of content analysis, Krippendorff's α (alpha) (e.g., Krippendorff 2013:Ch. 12), be used to assess the performance of different measurement strategies. Among the key properties of this index are that it accommodates any metric (including nominal, ordinal, interval, and ratio) and that, unlike many other reliability indices, it models "chance agreement" in a way that is both theoretically well-founded and pragmatically fruitful. As long as the researcher has access to a subsample that has been coded using both the new occupational classification and its predecessor—and such double-coded data are often available—it is possible to estimate the index and use it to subject different measurement strategies to systematic empirical evaluation.

We also conduct a case study in which we use the index to examine the effects of the switch from the 1970 to the 1980 Census Bureau Classification of Occupations in the United States. Unlike other changes in occupational classifications in the country, that switch involved a deep conceptual and empirical discontinuity (Vines and Priebe 1989). Researchers have interpreted this discontinuity as making very difficult, if not simply precluding, consistent measurement of social class across the 1970-1980 divide, and this

has greatly hindered the study of class-related trends. In the case of intergenerational mobility, for instance, Beller maintained that "[t]rends in social class fluidity after the mid-1980s are unclear, in part because changes to the census coding of occupations in the 1980s made it impossible to directly compare new survey data with older data" (Beller 2009:509; references omitted). It has been further argued that the problem is particularly serious when working with the widely-used Erikson–Goldthorpe–Portocarero (EGP) class scheme (e.g., Erikson and Goldthorpe 1992). For instance, in a study of social class differences in the earnings of black and white men, Morgan and McKerrow asserted that it is not possible to reconcile the 1970 and 1980 classifications without introducing substantial distortions into the data, especially when implementing the EGP class scheme (Morgan and McKerrow 2004, Supp. App.: 1), and for this reason they focused on the period starting in 1983.[2] The reported difficulties have not completely precluded the study of U.S. class-based trends across the 1970-1980 divide using the EGP scheme (see Hout 2005; Mitnik, Cumberworth and Grusky 2016; Pfeffer and Hertel 2015; Weeden et al. 2007). However, none of these studies provided any systematic assessment of the reliability of the class measures they used, either in absolute terms or compared to other possible measures.[3]

In our case study, we examine whether the obstacles to obtaining a reliable EGP measure of social class across the 1970-1980 classification divide are as fundamental as has been claimed. Using double-coded data from the General Social Survey (GSS; see Smith et al. 2011), we assess how seven different measurement strategies fare in terms of reliability and in terms of how they perform when used to estimate a common model of

intergenerational class mobility. Our results indicate that (a) the strategies, which seem equally reasonable a priori, vary substantially in terms of their reliability, (b) this variation in reliability is largely accounted for by the extent to which the strategies rely on researchers' subjective judgments about the relationships between occupational and class classifications, and (c) as long as the best-performing strategies are used, the problems generated by the 1970-1980 classification divide appear to be substantially less consequential than has been argued in the literature.

The paper is organized as follows. In the next section, we introduce Krippendorff's α and explain its conceptual advantages over several other widely-used reliability indices. The form of the index we present in this section is very general and makes its conceptual foundations transparent, but is not convenient for computational purposes. In the following section we therefore introduce a second expression—for nominal class variables only—that makes it much easier to compute point estimates and conduct statistical inference. The fourth section of the paper is devoted to our case study. The last section presents our main conclusions.

Reliability and social-class measurement strategies

In the context of measurement theory, reliability is the extent to which a measurement procedure provides the same results under repeated measurement of the same units; a procedure is considered reliable as long as it tends to produce consistent results across measurements. In turn, data are reliable as long as they have been produced using a reliable procedure. Reliability is distinguished from validity, which pertains to the

extent to which a measuring procedure measures what it purports to measure. Reliability is usually interpreted as a precondition for validity (e.g., Carmines and Zeller 1979).

When measuring social class across a change in occupational classifications, a researcher aims to use a reliable measurement procedure—that is, a procedure that assigns people the same class values regardless of which occupational classification was used to code their occupations. For each of the two occupational classifications, the researcher uses a set of rules to assign class values to occupations.[4] We refer to these sets of rules as "occupation-to-class mappings" or just "mappings". Together, the two mappings constitute a measurement strategy.[5] We can therefore say that the researcher's goal is to use a reliable measurement strategy.

If part of the sample (i.e., a subsample) has been doubled-coded using the two occupational classifications, the extent to which any strategy approaches the ideal of full reliability can be empirically assessed. To conduct such as assessment, the researcher would assign everyone in the double-coded sample two class values, using the two mappings in the measurement strategy. Loosely speaking, the greater the agreement between the class assignments across mappings, the more reliable the measurement strategy is.

In order to make this loose idea operational, the researcher needs a reliability index to quantify the degree of agreement, to determine if a particular strategy reaches the minimum level of reliability deemed acceptable, and to compare the reliability of different strategies. Below we critically examine a set of indices that have been used

extensively in various fields to assess reliability. Then we introduce Krippendorff's α, which is the index that we favor.

Critical assessment of widely-used reliability indices

An index that has often been used to assess the reliability of interval variables is the correlation coefficient. This is a very flawed approach, as it confuses predictability with reliability. In our context, the correlation coefficient would provide information about how well the class values produced by one occupation-to-class mapping linearly predict the values produced by the other. But one mapping's values can perfectly predict the other's even in cases where the two sets of values are completely different.

For nominal variables, the reliability index that perhaps seems most intuitive is the simple percent agreement—in our case, the percent agreement between two occupation-to-class mappings. For example, a measurement strategy where 80 percent of people are mapped into the same classes by both mappings seems clearly more reliable than a strategy where only 50 percent are mapped into the same classes. But intuitive as it may seem, percent agreement is a very unsound measure of reliability as well. It ignores that some level of agreement would occur even if class categories were assigned randomly.[6] Therefore it badly overstates the reliability of the data. Equally concerning, given that the agreement expected by chance, or "chance agreement," is a function of the number of class categories, using the percent agreement as reliability index makes meaningless any comparisons between class measures with different numbers of categories.

Several measures of reliability for nominal variables have been proposed that do take into account that agreement may occur by chance, and therefore include a correction for chance agreement. They all have the following form (Zwick 1988):

$$R = \frac{A_o - A_c}{1 - A_c}, \qquad [1]$$

where *R* is a reliability measure, A_c is the chance proportion of agreements as this is defined under *R*, and A_o denotes the observed proportion of agreements (i.e., the proportion of observations in which the mappings agree in their social class assignments).

Figure 1 shows the basic structure of a cross-tabulation of data used to assess the reliability of a measurement strategy for a nominal class variable. The figure shows the same *K* class categories in rows and columns. The class measure produced by mapping the first occupational classification into the class categories (mapping 1) is in rows, and the measure produced by mapping the second occupational classification into the class categories (mapping 2) is in columns. In the figure, p_{ij} denotes the proportion of people assigned to class *i* by mapping 1 and to class *j* by mapping 2, while $p_{i.}$ is the proportion of people assigned to class *i* by mapping 1 and $p_{.j}$ is the proportion of people assigned to class *j* by mapping 2. In all reliability measures for nominal variables, $A_0 = \sum_{k=1}^{K} p_{kk}$, that is, the observed agreement is simply the sum of the quantities in the main diagonal. The reliability measures differ, however, in how they define A_c.

Bennett, Alpert and Goldstein (1954) first proposed the reliability index *S*, which has since been reintroduced under many other names (see Zwick 1988). Under this index, chance is modeled as the random assignment of classes to people, assuming a uniform

probability over class categories. The probability of any ordered pair of values is then $\frac{1}{K^2}$, so the expected proportion of agreements is $A_c = K\frac{1}{K^2} = \frac{1}{K}$. The index equals one when there is perfect agreement and zero when agreement is as expected by chance. Its main shortcomings are that (a) the value of *S* can be large even if the two mappings produce very different marginal distributions of the class variable, when this should result in low reliability values; (b) if the mappings assign classes randomly, the values of *S* will be larger the more disagreement there is about marginal distributions (for these two issues, see Hsu and Field 2003); and (c) for a given value of A_o, the value of *S* increases as *K* (the number of classes) increases, even if no one is assigned the added classes (Scott 1955).

Cohen's (1960) κ (kappa) stipulates $A_c = \sum_{k=1}^{K} p_{k.}p_{.k}$. Thus, chance agreement here is the agreement that can be expected if classes are randomly and independently assigned to people within each mapping, with probabilities equal to the mapping-specific observed relative frequencies of the class categories. Although used widely in epidemiology and other fields, Cohen's κ is also seriously flawed. The main problem is that, for any given level of agreement (any value of A_o) κ *penalizes* agreements in marginal distributions across mappings, that is, κ is lower the more agreement there is between those distributions. The reason for this undesirable result is that the index takes the marginal distributions as priors; thus, two mappings that produce similar marginal distributions must achieve a much higher agreement rate to obtain a given value of κ than mappings that produce radically different marginal distributions (Brennan and Prediger 1981:692; see also Warrens 2012 and the references therein). Or, as Krippendoff (2004:

420-421) has shown, the better disagreement can be predicted, the higher the value of Kappa, which conflates predictability with agreement.

Scott's π (phi) defines $A_c = \sum_{k=1}^{K} \left(\frac{p_{k.}+p_{.k}}{2}\right)^2$, so chance agreement here is the agreement that can be expected if classes are assigned to people with probabilities equal to the best available estimates of their true prevalence in the population (Scott 1955:324), that is, their average prevalence across mappings. Unlike Cohen's κ, Scott's π does not confuse predictability with agreement (Krippendorff 2004:420-421). Although a sound reliability index, π has a "scope limitation" that it shares with S and κ: none of these indices can be used to establish the reliability of data based on metrics other than the nominal metric (e.g., ordinal, interval, ratio). As we indicate below, π is a particular case of our preferred and most general reliability index, Krippendorff's α.

Before introducing this index, a few words regarding Cronbach's α (Cronbach 1951), which is presented as a reliability measure and has often been used in sociology, seem in order. In our context, the relevant notion of reliability is reliability-as-replicability—the reproducibility of results across functionally equivalent measurement instruments. Chronbach's α, however, measures internal consistency (e.g., of the items in a psychometric test). Whatever its merits and limitations in this regard (see Sijtsma 2009 for a detailed analysis), it belongs to the family of correlation coefficients, and it shares their shortcomings if used as a measure of reliability.

Krippendorff's Alpha

Krippendorff's α is defined (e.g., Krippendorff 2013:Ch. 12) as

$$\alpha = 1 - \frac{D_o}{D_e},$$

where D_o and D_e are measures of observed and expected disagreement, respectively. Alpha ranges from 1 when there is no disagreement (that is, when there is perfect agreement), to 0 when observed and expected disagreement are the same.[7]

In our case, which involves two occupation-to-class mappings per measurement strategy, observed disagreement can be expressed as:

$$D_o(m) = \text{within person average difference}$$

$$= \frac{1}{n}\sum_{p=1}^{n} d_m\left(v_{1p}, v_{2p}\right),$$

where *n* is the number of people in the reliability sample (i.e., the double-coded sample); v_{qp} is the class assigned to person *p* by mapping q, $q = 1,2$; and d_m is the difference function under metric *m*, where *m* is nominal, ordinal, interval or ratio. Expressions for these difference functions are provided below.[8]

The disagreement expected by chance is here the expected disagreement when the class values assigned to people are statistically independent from people's actual values, and the marginal distribution of the values is the one implied by the actual assignments pooled across mappings. More precisely, assume a stochastic process in which the pair of class values found for each person (i.e., one value from each mapping) is the result of randomly selecting, with equal probability, one of the $2n(2n-1)$ 2-permutations (or partial permutations of size 2) of the set of $2n$ values actually assigned by the mappings to the *n* people in the reliability sample. A physical implementation of the random

assignment just described could be achieved as follows. For each unit in the reliability sample: (a) flip a coin to define which mapping "draws" first; (b) the selected mapping randomly draws one value from the *2n* available values; and (c) the other mapping randomly draws one value from the remaining (*2n* - 1) values. Under this model of chance, expected disagreement is the average difference across all possible $2n(2n-1)$ 2-permutations, which can be expressed as:

$$D_e(m) = \text{within data average difference}$$

$$= \frac{1}{2n(2n-1)} \sum_{i=1,j=1}^{n} \sum_{k=1,l=1}^{2} d_m(v_{kj}, v_{li}) \,.^{9}$$

The index can then be written as:

$$\alpha(m) = 1 - \frac{2(2n-1)\sum_{i=1}^{n} d_m(v_{1i}, v_{2i})}{\sum_{i=1,j=1}^{n}\sum_{k=1,l=1}^{2} d_m(v_{kj}, v_{li})}, \qquad [2]$$

where:

$$d_{nominal}(c,k) = I(c \neq k)$$

$$d_{ordinal}(c,k) = \left(\sum_{g=c}^{g=k} F_g - \frac{F_c + F_k}{2}\right)^2$$

$$d_{interval}(c,k) = (c-k)^2$$

$$d_{ratio}(c,k) = \left(\frac{c-k}{c+k}\right)^2,$$

I is the indicator function (i.e., a function that equals 1 if its argument is true and 0 otherwise), and F_j is the total number of observations assigned to category *j* by the two mappings combined.[10]

At least three well-known indices are particular cases of Krippendorff's α when, as here, only two mappings are included in a measurement strategy (Hayes and Krippendorff 2007; Krippendorff 1970): α approaches Scott's π when the sample is large and the metric is nominal, and it is equal to Spearman's rank correlation coefficient and to Pearson's intraclass correlation coefficient when the metrics are ordinal and interval, respectively. In addition, by making $D_o = 1 - A_0$ and $D_e = 1 - A_c$, it is easy to see that, in the case of the nominal metric, α is also a particular case of Zwick's (1988) expression (see Equation 1).[11]

Krippendorff's α is highly attractive as a reliability measure. It avoids the category mistake of conflating predictability and agreement. It takes into account that agreement may occur by chance, and it assesses deviations from perfect reliability by the proportion of observed to expected disagreement. Its (typical) range of variation has a clear interpretation, with 1 indicating perfect agreement and 0 indicating the same level of agreement as would be expected by chance.[12] By modeling chance agreement as the agreement that would occur if classes were assigned to people independently from their actual classes, it avoids the drawbacks associated with other measures that also take into account that agreement may occur by chance. In particular, Krippendorff's α punishes dissimilarity (and rewards similarity) between mapping-specific marginal distributions, and is fully unaffected by the inclusion or exclusion of unused class categories. Further, α allows fair reliability comparisons between class measures of different granularity, and between class measures that operationalize different theoretical approaches. And it can be computed for literally any metric, which has the additional advantage of allowing

reliability comparisons even across metrics (e.g., of nominal versus interval social class variables). Lastly, α admits several other useful interpretations (Krippendorff 2013: Ch. 12), including that it measures the extent to which the proportion of the undesirable to total disagreements subtracts from perfect agreement, where desirable and undesirable disagreements (in the assignment of class values to people) are those between and within people, respectively, and total disagreement is the sum of these two types of disagreements.

Krippendorff's Alpha with nominal class variables: Computation and statistical inference

The expression for Krippendorff's α provided by Equation 2 is very general but is not convenient for computational purposes. Computationally simpler expressions are available, in particular when class variables have a nominal metric (Krippendorff 2011). As this is the relevant metric for our case study and for much research on class-based trends, in what follows we focus exclusively on the nominal-metric case.

The computationally simpler expression we use is based on a coincidence matrix, which is a square and symmetric values-by-values matrix. This matrix records each pair of class values assigned to the people in the reliability sample twice, with the values in the pair "switching order." For instance, if person *p* was assigned the pair of class values (v_{1p}, v_{2p}), both this pair and (v_{2p}, v_{1p}) are included in the matrix.

Figure 2 shows an example of how a coincidence matrix is constructed. In the example, the reliability sample includes six people and the class scheme comprises three class positions: manual worker (MW), non-manual worker (NMW), and professional or manager (PM). The left panel of the figure shows the reliability sample as a standard

dataset with people as observation units and two variables containing the class values assigned by the two mappings. The right panel shows the coincidence matrix constructed from these data (observe, in particular, that the total number of assigned values, i.e., 12, is the same in both cases). Figure 3 shows the notation we use to refer to the different quantities in a general coincidence matrix. Using this notation, the much-easier-to-compute expression for α with a nominal class variable is:

$$\alpha = 1 - \frac{D_o}{D_e} = \frac{A_o - A_c}{1 - A_c} = \frac{\frac{\sum_{k=1}^{K} 0_{kk}}{t} - \frac{\sum_{k=1}^{K} t_k(t_k-1)}{t\,(t-1)}}{1 - \frac{\sum_{k=1}^{K} t_k(t_k-1)}{t\,(t-1)}}$$

$$= \frac{(t-1)\sum_{k=1}^{K} o_{kk} - \sum_{k=1}^{K} t_k\,(t_k-1)}{t(t-1) - \sum_{k=1}^{K} t_k\,(t_k-1)}, \qquad [3]$$

where $t = \sum_{k=1}^{K} t_k$ is the total number of assigned values (or twice the number of people in the reliability sample); t_k is the absolute frequency of class value *k* in the marginal distribution of class values (or the total number of times the value *k* was assigned by the mappings); and o_{kk} is the *k* element in the main diagonal of the coincidence matrix (or twice the number of people both mappings coded as belonging to class *k*). As should be apparent, in contrast to Equation [2], computation of Krippendorff's α with Equation [3] is a very simple exercise once the coincidence matrix has been generated.

In addition to providing the basis for an easy-to-compute expression, working with a coincidence matrix makes it possible to conduct statistical inference without having to resort to resampling approaches, and to deal with complex sampling designs easily. To the best of our knowledge, neither of these two advantages has been previously discussed in the literature on Krippendorff's α.

Let's assume for now that both the main sample and the reliability sample are simple random samples (SRS)—the main sample a SRS of the population of ultimate interest, and the reliability sample a SRS of the main sample. A value of α computed from a reliability sample is an estimate of the parameter of interest (the population to which that parameter pertains is discussed below), and is therefore subject to sampling variability. But the asymptotic distribution of $\hat{\alpha}$ is unknown, so previous research has proposed basing statistical inference on the nonparametric bootstrap (Hayes and Krippendorff 2007; Krippendorff 2013:Chapter 12). This is a highly computer- and time-intensive approach (the recommended number of bootstrap samples is 20,000), and its implementation is somewhat complicated (see Krippendorff 2016 [2006]).

The alternative approach we propose here is to use the "delta method" (e.g., Oehlert 1992). This involves computing, first, estimates of o_{kk} and t_k for $k = 1, 2, \ldots, K$, which we denote by $\hat{o}_{kk}$ and $\hat{t}_k$, and of their variance-covariance matrix, which we denote by $\widehat{\boldsymbol{V}}$. With this information in hand, we can use that $\hat{\alpha} = g(\hat{o}_{11} \ldots \hat{o}_{KK}, \hat{t}_1 \ldots \hat{t}_K)$ and therefore $\widehat{Var}(\hat{\alpha}) = \widehat{Var}(g(\hat{o}_{11} \ldots \hat{o}_{KK}, \hat{t}_1 \ldots \hat{t}_K))$, which allows a standard application of the delta method, and then do statistical inference using standard asymptotic procedures.[13] Estimates of the elements of the coincidence matrix, and their variance-covariance matrix, can be generated very easily from data in standard form, regardless of the number of class categories.[14]

Let's now assume that while the reliability sample is a SRS of the main sample, the latter is not a SRS of the population of ultimate interest. Rather, the main sample has been drawn using a "complex survey design" (e.g., Heeringa, West and Berglund 2010;

Skinner, Holt and Smith 1989), such as a design with clustering, stratification, and uneven probabilities of selection. How should we compute the coincidence matrix in this context?

One possible answer is that the parameter α pertains to the main sample, i.e., that it characterizes the reliability of the data in the main sample regardless of the relationship of these data with the ultimate population of interest. Under this interpretation, the complex survey design used to draw the main sample should be ignored, $o_{kk}$and t_k should be estimated using "unweighted estimators," and $\widehat{\boldsymbol{V}}$ should be computed using standard "SRS formulas."

A different—and, in our view, methodologically superior—answer is as follows. As we are ultimately concerned with the effects of data unreliability on the description of trends and the estimation of trend-relevant models, it seems natural to give more weight in our reliability assessments to the reliability of those data that represent larger segments of the population of interest. So, if there is a social class that is very unreliably coded, and the people that are assigned that class by either mapping have large sampling weights in the main sample, we want this fact to be reflected in our estimate of reliability. Formally, the parameter α would pertain in this case not to the main-sample data but to the population data. That is, imagine that the full population of interest is coded using the two mappings. In this context, (a) α has a specific value for the population, which we can estimate using "weighted estimators" of o_{kk} and t_k, and (b) our uncertainty regarding the true value of α given our point-estimate $\hat{\alpha} = g(\hat{o}_{11} \ldots \hat{o}_{KK}, \hat{t}_1 \ldots \hat{t}_K)$ can be assessed by

computing $\widehat{\boldsymbol{V}}$ with a "complex-survey formula," i.e., a formula that does take into account the complex survey design used to draw the main sample (e.g., Skinner et al. 1989).[15]

Regardless of what is deemed as the parameter of interest, if the reliability sample is not a SRS of the main sample, this typically cannot be ignored when generating the coincidence matrix—or, equivalently, the estimates of o_{kk} and t_k—and the associated matrix $\widehat{\boldsymbol{V}}$. Further, when the reliability sample is not a SRS of the main sample *and* the main sample has been drawn using a complex survey design, then estimating the parameter α for the ultimate population of interest rather than for the main-sample data requires taking into account both the sampling design used to draw the main sample from the population of interest and that used to draw the reliability sample from the main sample.

As a companion to this article, we have made available a Stata program to compute Krippendorff's α when the class variable is nominal. The program uses the delta method for statistical inference and allows for the main features of a complex survey design to be taken into account. With this program, assessing the reliability of nominal class variables using the approach we have advanced here is a very simple endeavor.[16]

3. Case study: Bridging the 1970-1980 U.S. occupational classification divide

As we mentioned in the introduction, unlike other changes in the U.S. Census Bureau Classifications of Occupations (COCs), the switch from the 1970 to the 1980 classification involved a deep conceptual and empirical discontinuity. Indeed, the COCs changed little between 1940 and 1960, and even the 1970 classification—which increased the number of categories by almost 50 percent compared to the 1960 classification—is

organized along lines analogous to its predecessors. The 1980 and 1990 classifications, for their part, are almost identical. However, in spite of the fact that it only involved a 14 percent increase in the number of categories compared to the 1970 classification, the 1980 classification constituted a major departure from previous ones (Vines and Priebe 1989).

As we pointed out earlier, researchers have interpreted this discontinuity as making very difficult, if not simply precluding, consistent measurement of social class across the 1970-1980 divide, in particular with the widely-used EGP class scheme (e.g., Erikson and Goldthorpe 1992). Here we will examine whether the difficulties are as fundamental as has been claimed. To this end we use data from the GSS 1988-1990, in which all observations have been double-coded using both the 1970 and the 1980 COCs.[17]

We start by briefly describing the logic and categories of the full version of the EGP class scheme. Next, we characterize the various measurement strategies we consider. After that, we present the main results of our reliability analysis, both for the class of the respondent and for the class of his or her father, and document the correspondence between a strategy's reliability (as measured by Krippendorff's α) and its performance in an illustrative analysis in which we estimate a widely-used model of intergenerational mobility. We finish the section by discussing the results of our case study.

The EGP class scheme

Table 1 shows the full version of the EGP class scheme, with the different classes identified, as is conventional, by roman numerals (later we will refer to "collapsed versions," in which the class scheme is defined at lower granularities). The theoretical foundations of the scheme have been discussed in detail by Erikson and Goldthorpe (1992) and Goldthorpe (2000), among others; here we provide a brief overview.

The EGP class scheme is based, in part, on the distinction between those who own the means of production and those who do not and are employed by others. Classes IVa, IVb, and IVc are made up of small-scale owners, or the "petty bourgeoisie." Class IVa is the nonfarm petty bourgeoisie with employees, class IVb is the nonfarm petty bourgeoisie without employees, and class IVc includes all farmers. The other ten classes are comprised mostly of non-owners/employees, but for partially pragmatic (and not uncontroversial) reasons, large employers are included in one of those classes (see Erikson and Goldthorpe 1992:40-41).

Employees are further differentiated by distinguishing between those whose jobs are regulated by a "labor contract" and those whose jobs are regulated by a "service relationship" with the employer (Erikson and Goldthorpe 1992). Jobs tend to be set up as labor contracts when the work they involve is relatively easy to monitor and the skills and knowledge they require are easily found in the market. In such circumstances, payment is for discrete amounts of work and there is little attempt to secure a durable relationship. On the other hand, jobs tend to be set up as service relationships when the skills and expertise they require are harder to find in the market, or when monitoring work is more

difficult. Service relationships offer strong incentives for the employee to align his or her interests with the employer, such as pre-established salary increases over time, well-defined career opportunities, job security, and pension rights.

In the EGP scheme, classes I and II (professional, administrative, and managerial workers) include those people whose jobs best fit the service-relationship characterization, while classes VI and VII (skilled and unskilled manual workers) and IIIb ("lower-grade" routine non-manual occupations) comprise those people whose jobs best fit the labor-contract characterization. The two remaining classes are mixed or intermediary cases. People in class IIIa (e.g., clerks) hold jobs that don't require specialized skills but do involve some monitoring difficulties. Those in class V (skilled manual workers) hold jobs where monitoring is not a problem but some specialized skills are required.

Measurement strategies

We examine seven different measurement strategies, which implement three different measurement approaches. As discussed earlier, any measurement strategy involves two mappings from occupation to social class—one for the first occupational classification, and one for the second. We term the three different approaches to generating these two mappings the *direct* approach, the *internal indirect* approach, and the *external indirect* approach. The direct approach involves defining the two mappings independently of each other. The occupations in each occupational classification (and possibly other variables) are separately mapped into social class categories by assessing

their conceptual relationship to the class categories. This approach was used by Hout (2005) and Pfeffer and Hertel (2015).

In the internal indirect approach, one of the two occupational classifications is first mapped into the other classification, and then the second classification is mapped into class categories—so here one of the final mappings is the result of concatenating two intermediary mappings. This approach requires a "crosswalk" between the two occupational classifications based on analyses of double-coded data. A crosswalk shows how the people in each category of one classification are distributed among the categories of the other classification. For instance, one of the crosswalks used in our analyses shows that among people coded as "architects" under the 1970 classification, 88 percent are classified as "architects" under the 1980 classification and the remaining 12 percent are split between "civil engineers" (7 percent) and "marine engineers and naval architects" (5 percent). Whenever more than one possible destination occupation exists for a given origin occupation, the occupation that is the most likely destination based on its relative frequency among potential destination occupations is the one selected. This indirect approach was employed by Mitnik et al. (2016) and Weeden et al. (2007).

In the external indirect approach, each primary occupational classification is mapped into an "external" occupational classification (which may be the same or different for the two primary classifications). Then the external classification (or classifications) are mapped into EGP classes. In this case, both final mappings are the result of concatenating two intermediary mappings. As discussed in more detail below,

Beller (2009) used an external classification in order to code EGP classes, but without crossing the 1970-1980 divide.

Table 2 describes the measurement strategies we examine, and Table 3 lists the EGP classes that these strategies produce at three granularities: high, intermediate, and low. The first strategy we consider (strategy A) uses the direct approach. This strategy was used by Hout (2005) and is based on two separate occupation-to-class mappings, one from the 1970 COC and one from the 1980 COC. At its highest granularity, the resulting classification includes nine classes; of these, three are classes in the full version of the EGP scheme, three are aggregations of classes from that scheme, and the remaining three are the result of taking managers out of classes I and II to separate them from upper and lower professionals.

Strategies B through E all employ the internal indirect approach, using in all cases crosswalks generated by the U.S. Census Bureau (see Vines and Priebe 1989: Tables 1 and 2). We refer to these crosswalks as the 1970→1980 crosswalk (showing the 1980 COC distribution for each occupation in the 1970 COC) and the 1980→1970 crosswalk (defined analogously). Strategy B uses the 1970→1980 crosswalk to map the 1970 COC into the 1980 COC, and then applies Hout's mapping of the 1980 COC into EGP classes also used in strategy A. Strategy C proceeds in a similar way, but with the 1970 COC as the "bridge" classification: it uses the 1980→1970 crosswalk to map the 1980 COC into the 1970 COC, and then uses Hout's mapping of the 1970 COC into EGP classes. Strategy D is based on a mapping of the 1980 COC into EGP classes that was used in an auxiliary sensitivity analysis by Beller (2009: Ftn. 4); this mapping was also developed

by Michael Hout.[18] In Table 2 we refer to it as the Beller-Hout mapping. We combine it with the 1970→1980 crosswalk to map the 1970 COC into EGP classes. This measurement strategy was used by Mitnik et al. (2016). Table 3 shows that strategies B through D produce the same approximate EGP categories as strategy A.

Strategy E is based on the mapping from the 1980 COC into EGP classes used by Morgan and Tang (2007). We combine it with the 1970→1980 crosswalk in order to map the 1970 COC into classes. A similar strategy was used by Weeden et al. (2007). Table 3 shows that, at its highest granularity, this strategy produces almost all of the classes in the full EGP scheme—the only exception being that classes IVa and IVb are collapsed into one class.

Finally, we examine two measurement strategies based on the external indirect approach. In both, the external classification used is the International Standard Classification of Occupations (ISCO). The ISCO is widely employed in cross-national research, as it was developed to give researchers a way to consistently classify occupations across countries. It has been revised several times, so to date there are four versions: ISCO-58, ISCO-68, ISCO-88, and ISCO-08. The GSS recodes the 1970 COC into ISCO-68, and it recodes the 1980 COC into both ISCO-68 and ISCO-88.[19]

Ganzeboom and Treiman have developed mappings from ISCO-68 and ISCO-88 into EGP classes; we will refer to these widely-used mappings as GT-68 and GT-88 (see, e.g., Ganzeboom and Treiman 1996). In her research with GSS data, Beller (2009) used the GSS's 1980 COC→ISCO-88 recode and GT-88 to generate EGP classes, at an intermediate granularity, for the period 1994-2007.[20] We use a similar approach with

measurement strategies F and G. Strategy F defines one mapping by combining the GSS's 1980 COC→ISCO-88 recode and GT-88, and the other by combining the GSS's 1970 COC→ISCO-68 recode and GT-68. Strategy G is similar, but substitutes the GSS's 1980 COC→ISCO-68 recode and GT-68 to define the first of the two mappings. Table 3 shows that, at their highest granularity, both strategies produce nine classes: seven are classes from the full version of the EGP scheme, and the remaining two are aggregations of classes in that scheme (IIIa + IIIb, and IVa + IVb).[21]

In addition to assessing the reliability of all measurement strategies at the highest granularity available, we also assess it at the intermediate granularity used in most empirical research. We consider two different ways of collapsing the full EGP scheme into six classes, which only differ in how they treat routine non-manual employees, lower grade (IIIb) (see Table 3).[22] The first stresses the divide between manual and non-manual workers, with all routine non-manual workers grouped into one class (IIIa + IIIb). This way of collapsing EGP classes is very close to the widely-used seven-class version (or CASMIN version) of the EGP scheme (Erikson and Goldthorpe 1992)—the only difference being that we assign all people working in agriculture to the same class, while the CASMIN version keeps farmers (IVc) and agricultural workers (VIIb) as two different classes. [23] We refer to this granularity as the "intermediate (manual/non-manual) granularity."

The second six-class version of the EGP scheme we consider includes all nonfarm lower-skill/lower-wage workers in one class, regardless of whether they are manual or non-manual workers, by combining classes IIIb and VII (nonfarm semi- and unskilled

manual workers). Mitnik et al. (2016) collapsed the EGP scheme in this way because prioritizing income similarities across classes rather than the manual/non-manual divide was more appropriate for their goal of exploring the relationship between the post-1980 income-inequality take off and social-fluidity in the U.S.[24] We refer to this granularity as the "intermediate (income-based) granularity."

Finally, we assess the various measurement strategies at the minimum level of granularity that is possible, i.e., when there are only two classes. Given recent arguments on the centrality of the divide between professional and managers, and all other classes, for the study of trends (Mitnik et al. 2016), we focus here on a two-class version of the EGP scheme that only distinguishes classes I and II from all other classes.

Main reliability results

We present here the main results of our reliability analysis. As previously indicated, our analysis is based on the double-coded data in the 1988-1990 GSS. We constrain our sample to subjects ages 18-64, and estimate the reliability of EGP class measures for men and women separately, and for "all" (i.e., men and women pooled). We also estimate the reliability of measures of father's class, in two different ways.[25] In one case, we use the same EGP schemes that we use for men and women and drop all observations for which the father was absent from the household when the subject was 16 years old (the information on the subject's father pertains to that age). In the other case, we proceed as Mitnik et al. (2016), and add the category "non-resident father" to the EGP scheme. We refer to the resulting "population" as "fathers (expanded)." Table 4 shows the number of observations used to assess reliability in each case.[26]

Table 5 presents point estimates of Krippendorff's α (in bold), and the corresponding confidence intervals (in parentheses).[27] The first three columns include the results for the EGP class of all respondents combined, and of men and women separately. The last two columns show the results for fathers' EGP class. Estimation of α is quite precise, so most of our discussion focuses on the point estimates.

The estimated values of α exhibit a great deal of variability—they span the range 0.56 to 0.90, with a mean of 0.77 and a median of 0.78. They also show clear patterns across strategies. Some of these patterns are particularly apparent in Figure 4, where we use different colors to distinguish the values of α according to the approach of the corresponding measurement strategy, and display those values separately for all respondents, men, women, fathers, and fathers (expanded) at each granularity.

The most important of these patterns is the near-perfect ordering of the strategies—within population-granularity pairs—by approach, with those strategies using the internal indirect approach (B, C, D, and E) outperforming all others in nearly every case, and the strategy using the direct approach (A) never outperformed by (and mostly outperforming) those using the external indirect approach (F and G). For the intermediate (manual/non-manual) granularity (the granularity most often employed in empirical research), the seven strategies are perfectly ordered by approach, regardless of population. At the other three granularities the pattern is very similar, though in some cases strategy A outperforms strategy E.

What drives the ordering of the strategies by approach? To answer, we need to distinguish between two types of mappings between classifications. There are mappings

that are generated through a fully objective procedure, such as those based on the 1970→1980 and 1980→1970 crosswalks. The crosswalks result from a simple cross-tabulation of a double-coded dataset. Once the double-coded data on which a crosswalk is based are available, the associated mapping could be fully generated—at least in principle—by a computer program. Therefore, we may refer to mappings like this as "mechanical mappings." In contrast, there are mappings that rely in an essential way on a researcher's judgment regarding the relationship between the categories in two classifications, as is the case with all mappings from occupational classifications into EGP classes. Although these mappings are based on objective facts that clearly constrain which occupations may be reasonably mapped into which EGP classes, they involve an unavoidable element of discretionary subjective judgment. We may refer to these mappings as "non-mechanical mappings."

Our hypothesis is that, within granularity-population pairs, reliability performance is mostly accounted for by the number of non-mechanical mappings a measurement strategy relies upon. Or, in other words, that reliability performance is largely accounted for by the extent to which a strategy relies on subjective judgments regarding relationships between classifications.

The strategies based on the internal indirect approach (which we found to produce the most reliable class measures) use only one non-mechanical mapping—the mapping of one occupational classification into EGP classes. Here the second occupational classification is mapped into the first through a mechanical mapping, and then mapped into classes using the same non-mechanical mapping employed with the

first classification.[28] The direct approach, in contrast, involves two non-mechanical mappings, because each occupational classification is mapped separately into classes. The one strategy we consider that is based on the direct approach, strategy A, is consistently out-performed by the four strategies based on the internal indirect approach.

The strategies we examine that use the external indirect approach involve even more non-mechanical mappings: strategy F uses four while strategy G uses three. (Strategy F uses non-mechanical mappings of the two COCs into two different ISCOs, and then uses two non-mechanical mappings to map the ISCOs into EGP classes. Strategy G differs only in that it maps both COCs into the same ISCO, so it requires only one ISCO→EGP mapping.) Consistent with our hypothesis that reliability is closely related to the number of non-mechanical mappings employed, these two strategies are consistently the least reliable of our seven, and F performs substantially worse than strategy G. Indeed, strategy F has the lowest reliability of all strategies, for all populations. This is the case for all granularities at which this strategy can be implemented, but most notably at the high measurement granularity, where its performance is about two standard deviations below the mean.[29]

In addition to the ordering of strategies by approach, some granularity-related patterns are also apparent. Table 5 and Figure 4 show that, regardless of strategy, there is in nearly all cases a gain in reliability (and never a loss) when switching from the high granularity to the other granularities. As Krippendorff's α does take into account that chance agreement is more likely as the granularity of a measure falls, this pattern reflects a real reliability penalty attached to working with high-granularity measures. At the same

time, the reliability gains from switching from the high to the low granularity are only substantial for strategies E, G, and F, and are largest for the latter (which, as indicated, performs very poorly at the high granularity). Moreover, for the best-performing strategies (B, C and D), reliability at the intermediate granularities tends to be higher than at *both* the high and the low granularities, although only slightly so in most cases. For five of the seven strategies, we can also compare the values of α across the two intermediate granularities. Table 5 indicates that the class measure is less reliable at the intermediate (income-based) granularity than at the intermediate (manual/non-manual) granularity in the case of strategies C and E, while the differences are either nil or negligible in the other cases.

For the most part, there are only small differences in reliability across "generations" (men versus fathers). When non-resident fathers are included, the values of α for fathers are quite substantially larger than those for men (Figure 4 and Table 5), but this is only because there is always agreement across mappings when non-resident fathers are added as a separate category. When non-resident fathers are excluded from the analysis, the values of α for fathers tend to be slightly smaller than those for men in the case of strategies A and B, and slightly larger otherwise.

Gender does not seem to make any difference, i.e., Table 5 does not suggest any consistent gender advantage in reliability. The values of α tend to be slightly higher for women than for men in strategies A, C, E, and G, the other way around in strategies D and F, with gender differences varying in sign across granularities in the case of strategy B.

Table 6 reports the results of a linear regression of α on the number of non-mechanical mappings (entered as a nominal variable), granularity, gender, and generation. All estimates are consistent with the patterns identified in our qualitative analysis. In particular, the coefficients for two, three, and four non-mechanical mappings are all negative, substantial in magnitude, and statistically significant. The value of the adjusted R^2 indicates that about four-fifths of the observed variation in α is accounted by the explanatory factors included in the regression.[30]

We finish our main reliability analysis by presenting, in Table 7, the results of testing the null hypothesis that the value of α is not larger than 0.75 for each measurement strategy, at each granularity and for each population we consider. We use 0.75 as a threshold for an "acceptable" level of reliability; this value is in line with thresholds that have been used in several other related contexts.[31] If the null hypothesis is rejected, we can be reasonably confident that a measurement strategy produces class measures that are reliable enough to be employed in empirical analyses.

This table underscores again that the strategies based on the internal indirect approach (strategies B through E) produce class measures that are clearly superior in their reliability to strategies based on the other approaches. Strategies B and D perform particularly well. For the latter, the null hypothesis can be rejected for all granularities and populations; for the former, it can be rejected for all granularities and populations except one. For the strategies using the external indirect approach—F and G—the null cannot be rejected for any population, regardless of granularity (ignoring the expanded

scheme for fathers). Strategy A, which uses the direct approach, falls somewhere in between.

At the highest level of granularity—and again ignoring the expanded scheme for fathers—the null can only be rejected for the best-performing strategies (i.e., B, C and D). At the intermediate manual/non-manual granularity most often used in empirical research, the null is very clearly rejected for all populations when using strategies B, C, D and E (all of which are based on the indirect internal approach and use only one non-mechanical mapping), but not with the other strategies.

Illustrative analysis: The core model of social fluidity

The "core model of social fluidity," developed by Erikson and Goldthorpe (1992), is used widely in the field of intergenerational class mobility (see, e.g., Breen 2004). We estimate it here for men ages 31-64, using the double-coded data from the GSS 1988-1990 we employed in our previous analyses.[32] For each of our seven measurement strategies, we estimate the same model twice, once using the EGP class measure based on the 1970 COC and once using the class measure based on the 1980 COC. We use class measures at the intermediate (manual/non-manual) granularity. This illustrative analysis will allow us to show that the extent to which results differ within measurement strategies, i.e., across occupational classifications, is strongly related to their reliability performance as measured by Krippendorff's α.

The reliability estimates for the sample used in the analysis are presented in Table 8. The table includes estimates of α for the EGP class of men ages 31-64 and for the EGP class of their fathers (as well as the corresponding p-values for the null hypothesis that

$\alpha \leq 0.75$). In the context of content analysis, Krippendorff (2013: 326) has argued that "it is a serious mistake to average the reliabilities of the variables of a complex instrument and take this average as a measure of overall data reliability . . . [g]enerally, when variables are equally important to the research effort . . . the lowest α among them is the reliability of all data" (italics omitted). This suggests using the minimum of men's and their fathers' α (or "minimum α") as the relevant measure of reliability when comparing strategies. This is what we do below.

We estimate the "U.S. variant" of the core model of social fluidity (Erikson and Goldthorpe 1992:121-131, 318-319). As the model is very well known, we will not present it in any detail here. However, to facilitate understanding of the discussion that follows, we remind the reader that this is a log-linear, or multiplicative, model of the cells of a mobility table (i.e., a cross-tabulation of the subject's and his father's EGP class), and that the predicted quantities in the model are the expected frequencies in each cell of that table. We will focus on eight key parameters: parameters h1 and h2, which aim to capture the effects on relative-mobility patterns of barriers to movements across classes at different hierarchical levels; parameters i1 and i2, which reflect the tendency of sons to inherit the class position of their fathers; parameter s, which measures barriers to movements between agricultural and nonagricultural class positions; and parameters a1, a2, and ax, which measure affinity among class positions. We refer to these eight parameters as "model-specific parameters." As is standard in models of relative mobility based on mobility tables, the core model also includes other parameters aimed at

capturing the main effects of the distribution of individuals over origin and destination classes, and the scale of the table.[33]

Figures 5a and 5b display, for each measurement strategy, a graph with the estimates of the model-specific parameters with the two EGP measures.[34] The darker bars show the parameter estimates using the class measure based on the 1970 COC; the lighter bars show the estimates using the class measure based on the 1980 COC. The strategies are arranged in decreasing order of reliability (as measured by their minimum α)—that is, the first graph corresponds to the strategy with the highest reliability, the second graph to the strategy with the second-highest reliability, and so forth. The first two graphs are those for strategies D and B, whose minimum α is close to 0.8. For these strategies with relatively high reliability, the estimates based on the 1970 and 1980 COCs are extremely similar (although clearly not identical). At the other end of the spectrum, the graphs for strategies A, G and F, whose minimum α is in the 0.67-0.72 range, show large within-graph discrepancies in the estimates of the inheritance parameters i1 and i2 (other discrepancies are similar in magnitude to those found in the case of strategies D and B). Importantly, the differences between the estimated inheritance parameters are very consequential for the qualitative conclusions that can be drawn. For instance, if i1 is equal to one, that means that—once all other forces included in the model are considered— sons are not more likely to be found in the same class positions as their fathers. The null hypothesis that this is the case cannot be rejected if the confidence interval for the parameter includes one. With strategies A, G, and F, the confidence interval covers the value one with one occupational classification but not with the other.

Not rejecting the null would go against "one of the most secure findings of mobility research" (Erikson and Goldthorpe 1992: 125).

The graphs for strategies E and C, whose minimum α is 0.76 in both cases, indicate that the magnitude of discrepancies for these cases is in between those found for the other two groups. Thus, considered jointly, the estimates in Figures 5a and 5b reveal a clear correspondence between reliability levels and qualitative assessments of the consistency of the estimates across occupational classifications.

For some purposes, researchers may care about the operation of the full model rather than about specific parameters. In this context, one way of assessing how much the estimates of the model as a whole differ across occupational classifications is by defining loss functions whose arguments are the differences in predicted frequencies in the cells of the mobility table. These loss functions capture the net effect of differences in the estimates of all parameters at once. The expectation is that we will observe again a clear correspondence between reliability and the losses defined by these loss functions.

As table totals vary slightly across measurement strategies, it is more convenient to work with predicted relative frequencies than with predicted frequencies (which are the quantities directly predicted by the model). To carry out the analysis, we define four loss functions that vary in two dimensions: (a) whether they measure the distance between predicted values as absolute differences or as absolute proportional differences, and (b) whether they use the arithmetic mean or the median as summary measure of the observed differences.[35] The resulting loss functions are thus the mean and median absolute differences in predictions, and the mean and median absolute proportional differences in

predictions, across occupational classifications.[36] We will refer to the losses defined by these four loss functions as "prediction losses."

Figure 6 presents plots of the prediction losses against the minimum α. It includes one graph for each loss function, with the loss in the vertical axis and the minimum α in the horizontal axis. For all loss functions, there is a clear negative relationship between the values of α and the size of the losses, with the dots representing the most reliable strategies clustered in the right-bottom quadrant of the graphs, and those representing the other strategies appearing in the left-upper quadrant. Spearman rank correlations are negative and large in all cases, i.e., their absolute values are in the 0.71-0.79 range. The expectation of a clear correspondence between values of α and prediction losses is borne out by these results.

Figure 6 also indicates that for the best-performing strategy in terms of reliability (strategy D), the mean and median differences in predicted relative frequencies are smaller than 0.3 and 0.2 percentage points, respectively. The proportional differences are also small, around 10 percent. The other three strategies with reliability above 0.75 (B, C and E) do not do as well as D—but they do not do much worse either, especially in terms of the median loss functions, which are much more robust to outlier differences. Among the three strategies with lower reliability values—all of which tend to post much larger losses—strategy A performs much worse than F and G, in spite of a larger minimum α. The reason for this is that with strategy A the disagreements in coded EGP classes across the 1970 and 1980 COCs are disproportionally concentrated in one category, nonfarm self-employment (disagreements are much more evenly distributed with all other

strategies). This has a large impact on the estimates of the main effects of the distribution of individuals over origin and destination classes, which translates into much larger differences in predicted frequencies.[37]

Discussion

Our case study has made clear that measurement strategies that may seem equally sensible a priori differ markedly in the reliability of the class measures they generate. At the crucial intermediate (manual/non-manual) granularity used most often in empirical research, reliability, as measured by Krippendorff's α, is between 13 and 27 percent higher with the best-performing measurement strategy than with the worst-performing one (with the exact figure depending on the population). The reliability differential is even higher at the high granularity, where α is between 26 and 38 percent higher with the best performer.

Importantly, there seems to be a simple logic to the reliability differences observed within granularity-population pairs, as the variation in performance among the seven strategies is largely accounted by the number of non-mechanical mappings they employ. Thus, the analysis suggests that the fewer "researcher degrees of freedom" that go into defining a measurement strategy, the more reliable the data tend to be.

The case study has uncovered trade-offs between reliability and granularity, but these are somewhat complex. Although it is the case that the lower the granularity the higher the percentage of agreements across occupational classifications, it is *not* the case that, in general, the lower the granularity the higher the reliability of the data. For nearly all populations and measurement strategies, there are reliability penalties associated with

measuring at a high granularity compared to measuring at both the intermediate and low granularities. However, there is no consistent penalty from measuring at the intermediate granularities compared to measuring at the low granularity, as in this case reliability only increases consistently across populations for the worst performers, while it often falls or stays the same for the more reliable strategies. Lastly, all granularity penalties tend to be rather small with the best-performing strategies.

We examined two different ways of collapsing the full EGP scheme into an intermediate granularity. Mitnik et al. (2016) have argued that the standard way of collapsing EGP categories in empirical analyses—i.e. the CASMIN version of the EGP scheme, which is similar to our intermediate (manual/non-manual) granularity—is, for some purposes, conceptually less appealing than an approach that prioritizes income similarities across classes rather than the manual/non-manual divide. Our results indicate that there is a non-negligible reliability cost from using the income-based approach with two of the five strategies for which we could compare the approaches. (Neither of those two strategies is the one used by Mitnik et al. in their own research.)

Our illustrative analysis focused on the core model of social fluidity, and showed a clear correspondence between values of α and qualitative assessments of the consistency of the estimates of its model-specific parameters across occupational classifications. A quantitative, and more holistic, assessment based on prediction losses also indicated correspondence. It showed a clear separation of the measurement strategies into two groups, with those with reliability lower than 0.75 tending to post much larger losses, and with the associated rank correlations between prediction losses and values of

α all above 0.7 and some close to 0.8. These results provide reassurance that α works as expected.

At the same time, in the analysis of prediction losses, *within-group ranks* based on the minimum α do not match well the corresponding ranks based on losses. This is particularly the case for the lower-reliability group, in which strategy A performs substantially worse than strategies F and G despite its higher α. Other empirical analyses, whose results we did not include here to save space, exhibit similar patterns. This suggests that researchers should *not* use values of α to mechanically rank measurement strategies, but rather should use them to (a) discard strategies that are below some minimum value deemed acceptable, (b) among those strategies that are acceptable, only use values of α to rank strategies (in terms of their reliability) if the differences between those values are substantial, and (c) report the value of α for the measures actually used in their empirical analyses.

What may be deemed an acceptable value of α? Based on threshold values used in related contexts (see note 31) and on the results of our empirical analyses, we suggest that, as rule of thumb, a class measure be judged acceptable if the null hypothesis that $\alpha \leq 0.75$ is rejected at the conventional significance level. And that, when comparing strategies for which this null hypothesis can be rejected, only differences in the value of α of about 0.04 or larger be considered informative. In spite of the fact that our acceptability criterion may seem somewhat conservative, the class variables generated by four of the measurement strategies we examined satisfy it for all populations (Table 7). In the case of the data used in our illustrative study (men ages 31-64 and their fathers),

Table 8 indicates that strategies B and D satisfy the criterion for both the subject- and the father-class variables (with strategies C and E satisfying it for the father but not for the subject).

4. Conclusions

Periodic changes in occupational classifications may be a serious obstacle for conducting research on class-based trends. The magnitude of the problem, however, does not only depend on the nature and scope of the changes, but also on the measurement strategies used to address them. We have proposed here that scholars interested in conducting research on class-based trends use Krippendorff's α to assess the reliability of class measures generated from data coded using different occupational classifications.

This reliability index has many desirable properties. It does not conflate predictability with agreement in class values. It takes into account that agreement may occur by chance, using for this purpose a methodologically appealing conception of chance. It has a range of variation with clear reliability interpretations. It may be computed regardless of the metric of the class variable. It may be legitimately used to compare measurement strategies across class schemes, metrics, and granularities, which allows evaluation of various reliability-related trade-offs (for instance, from working at a higher rather than at a lower granularity, or from using one class scheme rather than another). Lastly, coincidence-matrix formulas typically used to simplify computation also allow straightforward estimation and statistical inference—based not in the resampling approaches employed in the reliability literature but in the much less time- and computer-intensive delta method—even in the context of complex sampling designs. For nominal

variables, a computer program made available as a companion to the paper, and which implements the approach for statistical inference suggested here, makes both estimation and inference very simple and fast endeavors.

We put Krippendorff's α to work in conducting a case study of the effects of the switch from the 1970 to the 1980 U.S. Census Bureau occupational classifications on the reliability of EGP class measures. We found that the reliability of EGP class measures varies substantially across measurement strategies that, a priori, seem equally sensible. In addition, our analysis suggested that the main driving force behind that variation is the extent to which researchers' subjective judgments play a role in defining a strategy.

Class scholars have maintained that the discontinuity resulting from the switch in occupation classification has made it very difficult, if not impossible, to consistently measure social class across the 1970-1980 divide, in particular with the EGP class scheme. The results of our study suggest, however, that as long as the right measurement strategies are employed, the problems generated by that switch are substantially less consequential than previously argued. The best-performing strategies appear to produce data reliable enough to study class-based trends, in particular at the intermediary level of granularity typically used in empirical research.

Notes

[1] See for instance Wright and Martin (1987) on changes over time in the class structure, McCall and Manza (2011) on political attitudes across classes, and Breen and Jonsson (2007) on intergenerational social mobility.

[2] Other references to the difficulties generated by the 1970s-1980s divide can be found, for example, in Hauser (1998:14), Morgan and Cha (2007, Supp. App.:2), and Weeden and Grusky (2004).

[3] Pfeffer and Herter (2015: Note 7 and Online Supplement) report that they conducted a sensitivity analysis of the stability of the fluidity trends they find to changes in occupational classifications. Mitnik et al. (2016: Note 13) report testing the procedures they used to address the change in classifications by double coding EGP class using the 1970 and 1980 classifications in the three years (1988–1990) in which that is possible, and that the results were very consistent and no worse than those obtained using other procedures. Neither article offered any details about these analyses.

[4] As we mentioned earlier, sometimes other variables, such as indicators for self-employment or supervisory responsibilities, are used in addition to occupation in order to determine the class of a person. These other variables should always be assumed to be implicit when we refer to the derivation of classes from occupations, i.e., classes may be actually derived from occupations and other variables even when we do not mention the latter explicitly.

[5] To be more precise, the mappings are implicitly or explicitly *defined* by those sets of rules. A mapping is technically a mathematical function. It relates each occupation—that

is, each element in its domain—to one, and only one, class value—an element in its codomain. Thus, different sets of rules (which are typically implemented through computer code) may define the same mapping. If classes are derived from occupations and other variables, then the mapping is a multivariate function.

[6] As Mielke and Berry put it, reflecting the dominant position in the literature, "any agreement coefficient should reflect the amount of agreement in excess of what would be expected by chance" (2001:134).

[7] Although of no practical significance, as all reasonable measurement strategies should produce values of α well above zero, in principle the index can become negative if observed disagreement is higher than expected disagreement. This should only happen when there are small samples or systematic disagreements, e.g., if the mappings are designed with the goal of maximizing disagreements.

[8] Other metrics can also be used, but the four listed should cover all cases of interest for this paper. Although the index can be defined for situations in which any finite number of occupational classifications have been employed over time, we develop it here for the case that is relevant from a practical point of view, as subsamples in which observations have been coded with more than two occupational classifications are generally not available. For the extension to three or more classifications ("coders"), see Krippendorff (2011; 2013).

[9] This expression sums $4n^2$ terms, so the question may arise of why this is the average difference over $2n(2n-1)$ permutations. The trick is to recognize that $d_m(.) = 0$

whenever $i = j$ and $k = l$, regardless of the measure m. This happens $2n$ times, and $4n^2 - 2n = 2n(2n - 1)$.

[10] Assume there are 18 people in the sample and three classes: "high," "middle," and "low." If the mappings assign 10, 5 and 3 people, and 8, 4 and 6 people, respectively, to these three classes, we then have: $F_1 = 10 + 8 = 18$; $F_2 = 5 + 4 = 12$; and $F_3 = 3 + 6 = 9$.

[11] The formulation in terms of disagreements is necessary to deal with class measured at levels other than nominal.

[12] With respect to the lower bound of α, see the caveat in note 7.

[13] The function $g(.)$ is, of course, defined by Equation [3].

[14] This can be achieved with the help of a simple trick: (a) generate a copy of the reliability data, (b) switch the names of the mapping variables in this copy; (c) append the copy to the original reliability data; and (d) cross-tabulate the two mapping variables in the expanded dataset (thus producing the coincidence matrix), using any statistical software that is able to treat the elements in the cross-tabulation as estimates, to compute a cluster-corrected variance-covariance matrix (to account for the repeated observations in the expanded dataset), and to save that matrix for future use.

[15] The trick to generate the coincidence matrix described in the previous note can still be used in this context. The only restriction is that the statistical software employed needs to have full-fledged complex-survey commands, procedures or routines for doing cross-tabulations. Major statistical packages like Stata, SAS, and R meet this requirement.

[16] This Stata program, called “kanom,” was developed by the first author of this article and can be installed directly from within Stata, by using the following command when connected to the internet: “ssc install kanom.” Unlike the previously existing Stata program “krippalpha,” SAS and SPSS macros "kalpha" (Hayes and Krippendorff 2007), and R package “irr,” kanom allows taking into account the main features of complex survey designs (sampling weights, clusters, strata). Unlike “krippalpha” and “irr,” “kanom” produces confidence intervals and tests the null hypotheses that α is smaller or equal to 0.67, 0.75, and 0.8. Unlike the SAS and SPSS macros, which base statistical inference on the nonparametric bootstrap, in kanom inference is based on the delta method, as discussed in this article. The previously existing programs, however, allow computation of Krippendorff’s α with any of the four metrics introduced earlier, not just the nominal metric.

[17] Prior to 1988 the GSS only used the 1970 COC, while in 1991-2010 it only used the 1980 COC. In 2012, it switched to the 2010 COC.

[18] Personal communication from Emily Beller.

[19] To this end, the GSS uses mappings from the COCs into the ISCOs developed by Harry Ganzeboom (see Appendix I of the GSS codebook, available at gss.norc.org/Get-Documentation).

[20] The GT-88 mapping requires information on whether a worker is the supervisor of other workers. Since the GSS does not contain this information, Beller applied GT-88 under the assumption that all respondents were non-supervisors (personal communication from Emily Beller).

[21] GT-88 and GT-68 do distinguish between IVa and IVb, but in strategies F and G many cases belonging to IVa get coded as IVb due to the fact that these strategies assume supervisory status to be non-supervisor for everyone (see previous note). For this reason, we collapse these two categories into one.

[22] In addition to the two approaches we examine, still other approaches to measuring EGP classes at an intermediate granularity have been used (see, e.g., Pfeffer and Hertel 2015).

[23] Our decision to combine IVc and VIIb into one class is mostly pragmatically motivated. First, in contemporary advanced economies there are too few people working in agriculture, so they hardly show up in the available survey samples. Second, the mappings from occupations into EGP classes used in measurement strategies A to D — i.e., those employed by Hout (2005) and Beller (2009) —do not distinguish between classes IVc and VIIb (most likely, for the reason just mentioned).

[24] Mitnik et al. (2016) reported that mean income in class IIIb is much more similar to that of class VII than to that of class IIIa.

[25] We do not include mother's class in our analysis because mother's occupation only became available in the GSS in 1994.

[26] The main reason for the differences in sample size across measurement strategies is that the strategies differ in the precedency given to employment status and some occupations in assigning some classes to people. Thus, if employment status (but not occupation) or occupation (but not employment status) are missing, class may be missing under one strategy but non-missing under another (in fact, it may even be missing under one of a strategy's mappings and non-missing under the other, in which case it is dropped

from the analysis). An alternative way of proceeding is to consider "class missing" as a class category (whenever age and gender are not missing). With this approach, the sample sizes do not vary across measurement strategies and are 3,510 (both for all and for fathers), 1,584 (men) and 1,926 (women). We don't report the full results using this approach, but see note 30 for a summary.

[27] The estimates shown in Table 5 were obtained using sampling weights, after adjusting the weights so that all three years (1988-1990) in the reliability sample have the same influence on final results. Unweighted estimates are extremely similar. If we consider the GSS sample for 1972-2010 (which is the sample that researchers would most likely use in class-based trend analyses) as the main sample, then it is clear that the reliability sample is not a random subsample of the main sample. The reliability sample is identical to the main sample for 1988-1990, while observations in the main sample in previous and later years had a zero probability of being selected to the reliability sample. Therefore, strictly speaking we are estimating reliability for 1988-1990—and interpreting the results as estimates of the reliability for 1972-2010 under the additional (untestable) assumption that reliability is the same across years.

[28] For instance, strategy D resorts to a mapping based on the 1970→1980 crosswalk and to the Beller-Hout mapping of the 1980 COC into EGP classes, but only the latter is non-mechanical.

[29] The exact values (i.e., standard deviations below the mean) are 2.2 (all), 2.6 (men), 2.5 (women), 1.65 (fathers) and 0.84 (fathers, expanded).

[30] Although the values of α are all higher, the relative-reliability patterns described in this subsection remain essentially unchanged if we code class missing as a category (see note 26).

[31] Cronbach's α is widely employed to assess the internal consistency of responses in multi-item scales, and while 0.7 is often considered the threshold for using a scale it has been argued that "increasing reliabilities much beyond 0.8 in basic research is often wasteful of time and money" (e.g., Nunnally and Bernstein 1994:264-265). With several measures used to assess the reliability of psychopathology diagnoses in psychiatry (including Cohen's κ), values between 0.61 and 0.8 are traditionally assumed to indicate "moderate" reliability, while values above 0.8 to indicate "substantial" reliability (Shrout 1998:308). In a well-known statistical textbook, Fleiss and his coauthors wrote that, for most purposes, values of Cohen's κ "larger than 0.75 or so may be taken to represent excellent agreement beyond chance" (Fleiss, Levin and Paik 2003:609). In content analysis it is customary to use 0.8 as the minimum value of α required to consider the data reliable, while values between 0.67 and 0.8 are deemed to characterize data that only should be used to draw tentative conclusions.

[32] We limit the analysis to this age group because younger people are arguably too young for their reported class position to offer a good representation of their long-term destination class.

[33] We estimate the core model of social fluidity with a six-class version of the EGP scheme rather than the seven-class version used in the formulation of the model (as mentioned earlier, the six-class version we use collapses farmers and workers in the farm

sector into one class). The specification of the model has been slightly adjusted to account for this fact.

[34] The estimates pertain to the parameters in the multiplicative specification of the model. The parameters in this specification are identical to the exponential of the parameters in the log-linear specification of the model (and are therefore always positive).

[35] Using a summary measure to define a loss function is necessary because there are 36 differences—one per cell of the mobility table—that need to be mapped into just one number quantifying the loss associated to the corresponding measurement strategy.

[36] Each absolute proportional difference is computed by dividing the absolute difference between predicted values by the mean predicted value. For instance, if, with a given strategy, the predicted value for the cell of the table where father and son are both professional or managers is 0.2 when using the 1970 COC to assign EGP classes, while the predicted value using the 1980 COC is 0.3, the absolute proportional difference is:

$$\frac{|0.2 - 0.3|}{(0.2 + 0.3)/2} = \frac{0.1}{0.25} = 0.4,$$

or, with a slight abuse of language, “40 percent.”

[37] The proportional absolute differences in estimates (across the 1970 and 1980 COCs) for the main effect of nonfarm self-employment are the following:

	MS A	MS F	MS G
Father	36.5 %	12.7 %	13.7 %
Son	19.0 %	1.3 %	4.0 %

References

Beller, Emiliy. 2009. "Bringing Intergenerational Social Mobility Research into the Twenty-first Century: Why Mothers Matter." *American Sociological Review* 74(4):507-528.

Bennett, E., R. Alpert, and A Goldstein. 1954. "Communications Through Limited-Response Questioning." *Public Opinion Quarterly* 18(3):303-308.

Breen, Richard (Ed.). 2004. *Social Mobility in Europe*. Oxford: Oxford University Press.

Brennan, R. L., and D. J. Prediger. 1981. "Coefficient Kappa: Some Uses, Misuses, and Alternatives." *Educational and Psychological Measurement* 41: 687–699.

Carmines, Edward, and Richard Zeller. 1979. *Reliability and Validity Assessment*. London: Sage.

Cohen, J. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20:37-46.

Cronbach, L. J. 1951. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrica* 16:297-334.

Erikson, Robert, and John Goldthorpe. 1992. *The Constant Flux: A Study of Class Mobility in Industrial Societies*. New York: Oxford University Press.

Fleiss, Joseph L. , Bruce Levin, and Myunghee Cho Paik. 2003. *Statistical Methods for Rates and Proportions*. New York: Wiley.

Ganzeboom, Harry, and Donald Treiman. 1996. "Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations." *Social Science Research* 25: 201-239.

Goldthorpe, John. 2000. *On Sociology: Numbers, Narratives and the Integration of Research and Theory*. Oxford: Oxford University Press.

Hauser, Robert. 1998. "Intergenerational Economic Mobility in the United States. Measures, Differentials and Trends." CDE Working Paper No 98-12, Center for Demography and Ecology, University of Wisconsin-Madison.

Hayes, Andrew, and Klaus Krippendorff. 2007. "Answering the Call for a Standard Reliability Measure for Coding Data." *Communication Methods and Measures* 1(1):77-89.

Heeringa, S. G., B. T. West, and P. A. Berglund. 2010. *Applied Survey Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Hout, Michael. 2005. "Educational Progress for African-Americans and Latinos in the United States from the 1950s to the 1990s: The Interaction of Ancestry and Class." in *Ethnicity, Social mobility and Public Policy: Comparing the USA and UK*, edited by Glenn Loury, Tariq Modood, and Steven Michael Teles. Cambridge: Cambridge University Press.

Hsu, Louis, and Ronald Field. 2003. "Interrater Agreement Measures: Comments on Kappa_n, Cohen's Kappa, Scott's π, and Aickin's α." *Understanding Statistics* 2(3):205-219.

Krippendorff, Klaus. 1970. "Bivariate Agreement Coefficients for Reliability of Data." *Sociological Methodology* 2:139-150.

—. 2004. "Reliability in Content Analysis. Some Common Misconceptions and Recommendations." *Human Communication Research* 30(3):411-433.

—. 2011. "Computing Krippendorff's Alpha-Reliability." Available at web.asc.upenn.edu/usr/krippendorff/mwebreliability5.pdf.

—. 2013. *Content Analysis: An Introduction to its Methodology (3rd ed.)*. Los Angeles: Sage

—. 2016 [2006]. "Bootstrapping Distributions for Krippendorff's Alpha ". Available at web.asc.upenn.edu/usr/krippendorff/boot.c-Alpha.pdf.

Mielke, Paul, and Kenneth Berry. 2001. *Permutation Methods: A Distance Function Approach*. New York: Springer.

Mitnik, Pablo, Erin Cumberworth, and David Grusky. 2016. "Social Mobility in a High-Inequality Regime." *The Annals of the American Academy of Political and Social Science* 663(1):140-184.
Morgan, Stephen, and Youngjoo Cha. 2007. "Rent and the Evolution of Inequality in Late Industrial United States." *American Behavioral Scientist* 50(5):677-701.
Morgan, Stephen, and Mark McKerrow. 2004. "Social Class, Rent Destruction, and the Earnings of Black and White Men, 1982-2000." *Research in Social Stratification and Mobility* 21:215-251.
Morgan, Stephen, and Zun Tang. 2007. "Social Class and Workers' Rent, 1983-2001." *Research on Social Stratification and Mobility* 25:273-293.
Nunnally, J. C. , and I. H. Bernstein. 1994. *Psychometric Theory*. New York: McGraw-Hill.
Oehlert, Gary. 1992. "A Note on the Delta Method." *The American Statistician* 46(1):27-29.
Pfeffer, Fabian, and Florian Hertel. 2015. "How Has Educational Expansion Shaped Social Mobility Trends in the United States?" *Social Forces* 94(1):143-180.
Scott, William. 1955. "Reliability of Content Analysis. The Case of Nominal Scale Coding." *Public Opinion Quarterly* 19:321-325.
Shrout, Patrick. 1998. "Measurement Reliability and Agreement in Psychiatry." *Statistical Methods in Medical Research* 7:301-317.
Sijtsma, Klaas. 2009. "On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha." *Psychometrica* 74(1):107-120.
Skinner, C. J., D. Holt, and T. M. F. Smith (Eds.). 1989. *Analysis of Complex Surveys*. New York: Wiley.
Smith, Tom W., Peter H. Marsden, Michael Hout, and Kim Jibum. 2011. "General Social Surveys, 1972-2010 [machine-readable data file]." Chicago: National Opinion Research Center.
Vines, Paula L., and John A. Priebe. 1989. *The Relationship Between the 1970 and 1980 Industry and Occupation Classification Systems*. Washington D.C.: U.S. Bureau of the Census.
Warrens, Matthijs. 2012. "The Effects of Combining Categories on Bennett, Alpert and Goldstein's S." *Statistical Methodology* 9:341-352.
Weeden, Kim, and David Grusky. 2004. "Are There Any Big Classes at All?" *Research in Social Stratification and Mobility* 22:3-56.
Weeden, Kim, Young-Mi Kim, Matthew Di Carlo, and David Grusky. 2007. "Social Class and Earnings Inequality." *American Behavioral Scientist* 50:702-736.
Zwick, Rebecca. 1988. "Another Look at the Inter-Rater Agreement." *Psychological Bulletin* 103(3):374–387.

Table 1: EGP class scheme, full version

I	Higher-grade professionals, administrators, and officials; managers in large industrial establishments; large proprietors
II	Lower-grade professionals, administrators and officials; higher-grade technicians; managers in small industrial establishments; supervisors of non-manual employees
IIIa	Routine non-manual employees, higher grade (administration and commerce)
IIIb	Routine non-manual employees, lower grade (sales and services)
IVa	Small proprietors, artisans, etc., with employees
IVb	Small proprietors, artisans, etc., without employees
IVc	Farmers and smallholders; other self-employed workers in primary production
V	Lower-grade technicians; supervisors of manual workers
VI	Skilled manual workers
VIIa	Semi-and unskilled manual workers (not in agriculture, etc.)
VIIb	Agricultural and other workers in primary production

Table 2: Measurement strategies

Measurement strategy	Approach	Description
A	Direct	Hout's (2005) mappings of the 1970 and 1980 COCs into EGP classes
B	Indirect, internal	Hout's (2005) mapping of the 1980 COC into EGP classes and 1970→1980 crosswalk
C	Indirect, internal	Hout's (2005) mapping of the 1970 COC into EGP classes and 1980→1970 crosswalk
D	Indirect, internal	Beller-Hout mapping of the 1980 COC into EGP classes and 1970→1980 crosswalk
E	Indirect, internal	Morgan and Tang's (2007) mapping of the 1980 COC into EGP classes and 1970→1980 crosswalk
F	Indirect, external	GSS's 1980 COC→ISCO-88 and 1970 COC→ISCO-68 recodes, plus GT-88 and GT-68 mappings into EGP classes
G	Indirect, external	GSS's 1980 COC→ISCO-68 and 1970 COC→ISCO-68 recodes, plus GT-68 mapping into EGP classes

Table 3: EGP classes at different granularities

Granularity	Measurement strategy		
	A, B, C and D	E	F and G
Highest available	I excp. managers II excp. managers Managers IIIa IIIb IVa + IVb V + VI VIIa IVc + VIIb	I II IIIa IIIb IVa + IVb IVc V VI VIIa VIIb	I II IIIa + IIIb IVa + IVb IVc V VI VIIa VIIb
Intermediate (stressing manual / nonmanual divide)	I + II IIIa + IIIb IVa + IVb V + VI VIIa IVc + VIIb	I + II IIIa + IIIb IVa + IVb V + VI VIIa IVc + VIIb	I + II IIIa + IIIb IVa + IVb V + VI VIIa IVc + VIIb
Intermediate (stressing similarity in income)	I + II IIIa IVa + IVb V + VI VIIa + IIIb IVc + VIIb	I + II IIIa IVa + IVb V + VI VIIa + IIIb IVc + VIIb	
Low (Professional/Managers versus all other classes)	I + II All other classes	I + II All other classes	I + II All other classes

Table 4: Number of observations with non-missing class values

Measurement strategy	Population				
	All	Men	Women	Fathers	Fathers (exp.)
A	3,314	1,530	1,784	2,904	3,389
B	3,293	1,510	1,783	2,858	3,343
C	3,316	1,532	1,784	2,909	3,394
D	3,289	1,506	1,783	2,856	3,341
E	3,282	1,502	1,780	2,838	3,323
F	3,302	1,514	1,788	2,881	3,366
G	3,329	1,539	1,790	2,931	3,416

Table 5: Estimates of Krippendorff's α

Measurement strategy	Measurement granularity	Population				
		All	Men	Women	Fathers	Fathers (exp.)
A	High	**0.75** (0.73-0.77)	**0.73** (0.70-0.75)	**0.76** (0.73-0.78)	**0.73** (0.71-0.75)	**0.77** (0.76-0.79)
	Intermediate (manual/nonm.)	**0.77** (0.75-0.79)	**0.75** (0.72-0.78)	**0.78** (0.75-0.80)	**0.74** (0.72-0.76)	**0.78** (0.77-0.80)
	Intermediate (income)	**0.76** (0.74-0.78)	**0.74** (0.71-0.77)	**0.77** (0.74-0.80)	**0.74** (0.72-0.76)	**0.78** (0.77-0.80)
	Low	**0.77** (0.75-0.80)	**0.77** (0.73-0.80)	**0.77** (0.74-0.81)	**0.76** (0.73-0.79)	**0.84** (0.82-0.86)
B	High	**0.78** (0.77-0.80)	**0.78** (0.76-0.80)	**0.78** (0.75-0.80)	**0.78** (0.77-0.80)	**0.82** (0.80-0.83)
	Intermediate (manual/nonm.)	**0.81** (0.79-0.83)	**0.81** (0.78-0.83)	**0.80** (0.77-0.82)	**0.79** (0.77-0.81)	**0.83** (0.81-0.84)
	Intermediate (income)	**0.80** (0.78-0.81)	**0.80** (0.77-0.82)	**0.79** (0.76-0.81)	**0.79** (0.77-0.81)	**0.83** (0.81-0.84)
	Low	**0.79** (0.77-0.81)	**0.78** (0.75-0.82)	**0.80** (0.77-0.83)	**0.76** (0.73-0.78)	**0.84** (0.82-0.86)
C	High	**0.77** (0.75-0.78)	**0.75** (0.73-0.78)	**0.77** (0.74-0.79)	**0.79** (0.77-0.81)	**0.83** (0.81-0.84)
	Intermediate (manual/nonm.)	**0.81** (0.79-0.82)	**0.79** (0.76-0.82)	**0.81** (0.78-0.83)	**0.81** (0.79-0.83)	**0.84** (0.83-0.86)
	Intermediate (income)	**0.77** (0.75-0.79)	**0.77** (0.74-0.79)	**0.77** (0.74-0.80)	**0.80** (0.78-0.82)	**0.84** (0.82-0.85)
	Low	**0.79** (0.76-0.81)	**0.78** (0.74-0.81)	**0.79** (0.76-0.82)	**0.79** (0.77-0.82)	**0.86** (0.84-0.88)
D	High	**0.79** (0.78-0.81)	**0.80** (0.78-0.82)	**0.78** (0.76-0.81)	**0.83** (0.82-0.85)	**0.86** (0.84-0.87)
	Intermediate (manual/nonm.)	**0.82** (0.80-0.84)	**0.83** (0.80-0.85)	**0.80** (0.77-0.82)	**0.84** (0.82-0.86)	**0.87** (0.85-0.88)
	Intermediate (income)	**0.81** (0.79-0.82)	**0.82** (0.79-0.84)	**0.79** (0.76-0.81)	**0.84** (0.82-0.86)	**0.87** (0.85-0.88)
	Low	**0.81** (0.79-0.83)	**0.82** (0.78-0.85)	**0.80** (0.77-0.84)	**0.85** (0.82-0.87)	**0.90** (0.89-0.92)
E	High	**0.72** (0.70-0.74)	**0.70** (0.67-0.73)	**0.73** (0.70-0.75)	**0.75** (0.73-0.77)	**0.79** (0.77-0.81)
	Intermediate (manual/nonm.)	**0.80** (0.78-0.81)	**0.78** (0.76-0.81)	**0.80** (0.77-0.82)	**0.81** (0.80-0.83)	**0.85** (0.83-0.86)
	Intermediate (income)	**0.76** (0.74-0.77)	**0.75** (0.72-0.77)	**0.75** (0.73-0.78)	**0.80** (0.78-0.82)	**0.83** (0.82-0.85)
	Low	**0.77** (0.74-0.79)	**0.75** (0.71-0.79)	**0.78** (0.75-0.82)	**0.79** (0.77-0.82)	**0.86** (0.85-0.88)
F	High	**0.58** (0.56-0.60)	**0.58** (0.55-0.61)	**0.56** (0.53-0.59)	**0.62** (0.60-0.64)	**0.68** (0.66-0.70)
	Intermediate (manual/nonm.)	**0.66** (0.64-0.68)	**0.67** (0.64-0.70)	**0.63** (0.59-0.66)	**0.72** (0.70-0.74)	**0.77** (0.76-0.79)
	Low	**0.72** (0.70-0.75)	**0.72** (0.68-0.76)	**0.72** (0.69-0.76)	**0.72** (0.69-0.75)	**0.81** (0.79-0.83)
G	High	**0.69** (0.67-0.71)	**0.67** (0.64-0.70)	**0.69** (0.66-0.72)	**0.71** (0.69-0.73)	**0.76** (0.74-0.77)
	Intermediate (manual/nonm.)	**0.71** (0.69-0.73)	**0.69** (0.66-0.72)	**0.70** (0.67-0.73)	**0.73** (0.71-0.75)	**0.78** (0.76-0.80)
	Low	**0.75** (0.73-0.78)	**0.75** (0.71-0.79)	**0.76** (0.72-0.79)	**0.76** (0.73-0.78)	**0.84** (0.82-0.86)

Note: Point estimates in bold, 95 % confidence intervals in parentheses

Table 6: Linear regression of Krippendorff's α on number of non-mechanical mappings, granularity, gender and generation

	Krippendorff's α
Number of non-mechanical mappings	
Two	-0.04**
Three	-0.07**
Four	-0.12**
Granularity	
Intermediate (manual/nonmanual)	0.04**
Intermediate (income based)	0.03**
Low	0.05**
Gender	
All (men and women)	0.01
Women	0.00
Generation	
Father	0.02*
Father (expanded)	0.07**
Constant	0.75**
Observations	130
Adjusted R^2	0.78

Notes:

1. * $p < 0.05$, ** $p < 0.001$.
2. Omitted categories are One mapping (Number of non-mechanical mappings), High (Granularity), Men (Gender) and Subject (Generation).

Table 7: Hypothesis tests for Krippendorff's $\alpha \leq 0.75$

Measurement strategy	Measurement granularity	Population				
		All	Men	Women	Fathers	Fathers (exp.)
A	High					**
	Intermediate (M/NM)	*		*		***
	Intermediate (Inc.)					***
	Low	*				***
B	High	***	**	**	***	***
	Intermediate (M/NM)	***	***	***	***	***
	Intermediate (Inc.)	***	***	**	***	***
	Low	***	*	**		***
C	High	*			***	***
	Intermediate (M/NM)	***	**	***	***	***
	Intermediate (Inc.)	**			***	***
	Low	**		**	***	***
D	High	***	***	**	***	***
	Intermediate (M/NM)	***	***	***	***	***
	Intermediate (Inc.)	***	***	**	***	***
	Low	***	***	***	***	***
E	High					***
	Intermediate (M/NM)	***	**	***	***	***
	Intermediate (Inc.)				***	***
	Low			*	***	***
F	High					
	Intermediate (M/NM)					**
	Low					***
G	High					
	Intermediate (M/NM)					***
	Low					***

Note:
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

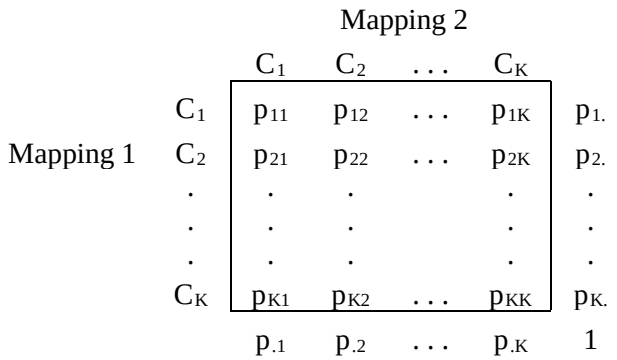
Table 8: Estimates of Krippendorff's α, core-model class variables (men ages 31-64)

Measurement Strategy	Number of observations	Men's α	p-value	Men's Fathers α	p-value
A	935	**0.72** (0.68-0.76)	0.934	**0.72** (0.68-0.75)	0.961
B	913	**0.79** (0.76-0.82)	0.011	**0.79** (0.76-0.82)	0.007
C	940	**0.76** (0.73-0.80)	0.221	**0.79** (0.76-0.82)	0.007
D	909	**0.81** (0.78-0.84)	0.000	**0.82** (0.79-0.85)	0.000
E	904	**0.76** (0.72-0.80)	0.308	**0.81** (0.78-0.84)	0.000
F	924	**0.67** (0.63-0.71)	1.000	**0.72** (0.69-0.76)	0.949
G	951	**0.68** (0.64-0.72)	1.000	**0.73** (0.70-0.76)	0.869

Notes:

1. Class measured at the intermediate (manual/nonmanual) granularity in all cases.
2. Point estimates in bold, 95 % confidence intervals in parentheses.
3. The p-values in columns 4 and 6 are for H0: $\alpha \leq 0.75$.

Figure 1: Cross-tabulation of nominal class measures produced by a measurement strategy's mappings



$C_1, C_2, \ldots, C_K$ are class categories

Figure 2: Example of reliability sample (standard and coincidence-matrix representations)

People-by-variables dataset

ID	Mapping 1	Mapping 2
1	MW	MW
2	MW	NMW
3	NMW	NMW
4	NMW	PM
5	PM	PM
6	PM	PM

Coincidence matrix

	MW	NMW	PM	
MW	2	1	0	3
NMW	1	2	1	4
PM	0	1	4	5
	3	4	5	12

MW: Manual worker; NMW: Non-manual worker; PM: Professional or Manager.

Figure 3: General coincidence matrix

	1	.	k	.	.	
1	O_{11}	.	O_{1k}	.	.	t_1
.	.		.			.
.	.					.
r	O_{r1}	.	O_{rk}	.	.	$t_r = \sum_k O_{rk}$
.	.		.			.
	t_1	.	t_k		.	$t = \sum_r \sum_k O_{rk}$

Figure 4: Krippendorff's α by granularity and approach

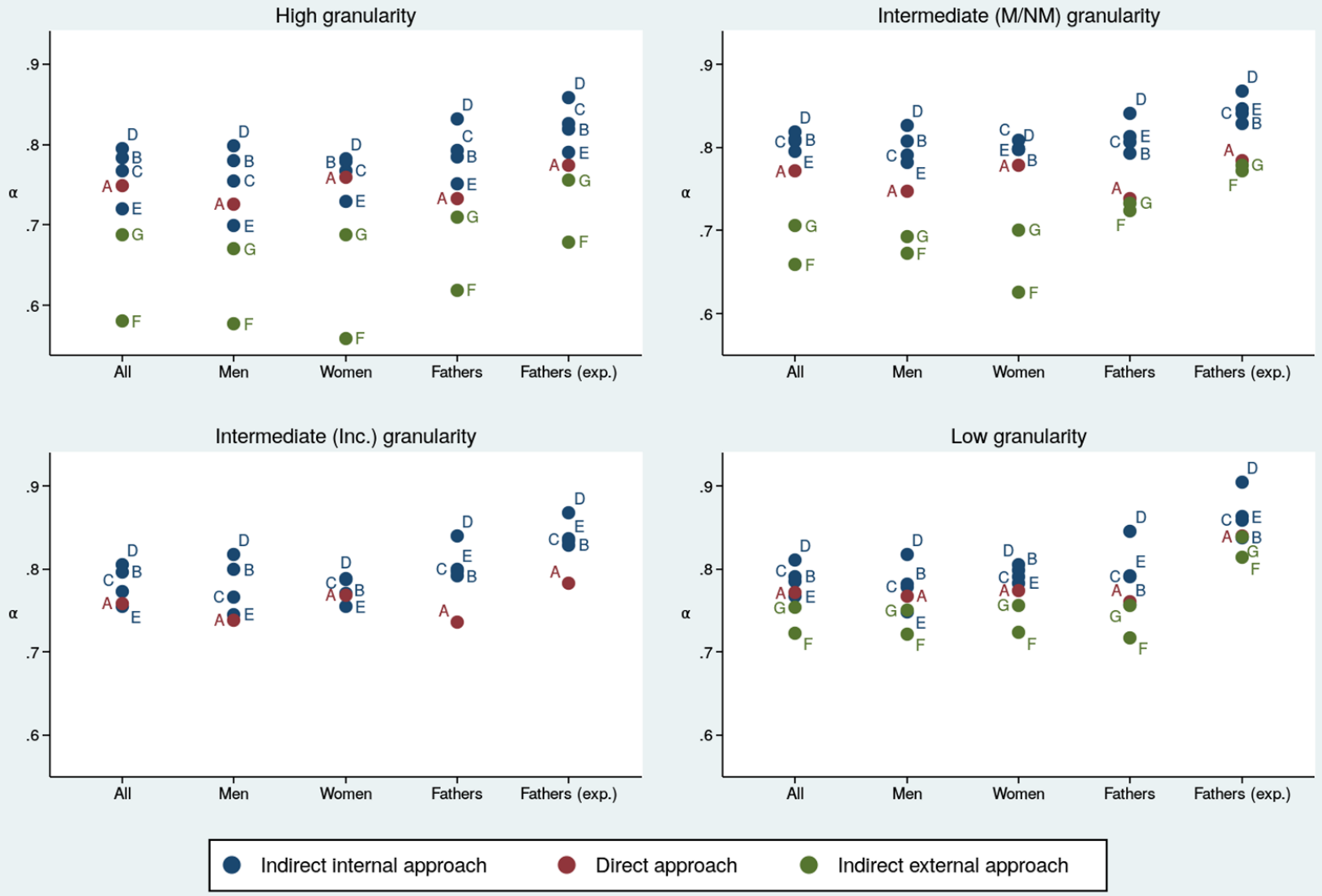


Figure 5a: Core-model parameter estimates (model-specific parameters)

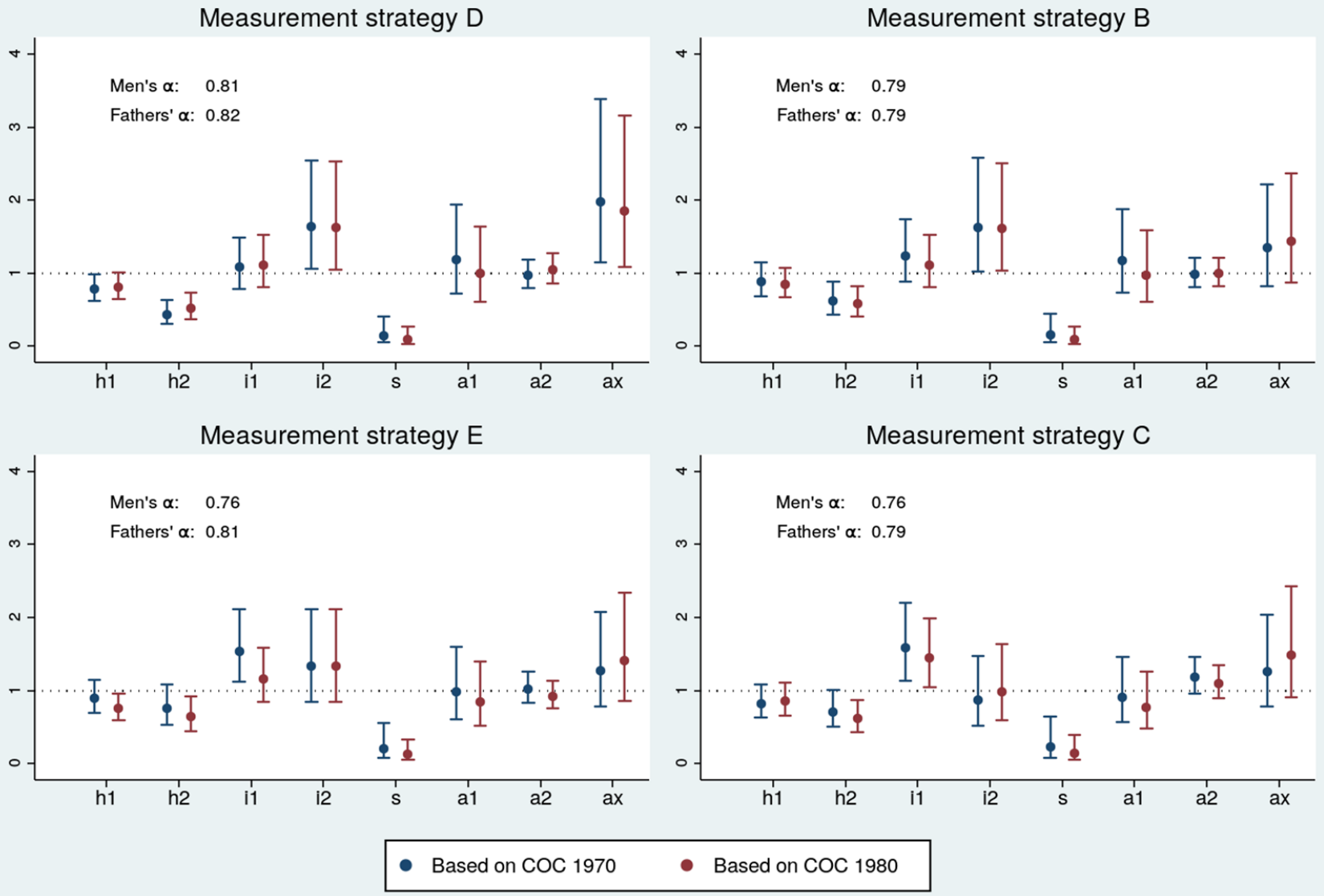


Figure 5b: Core-model parameter estimates (model-specific parameters)

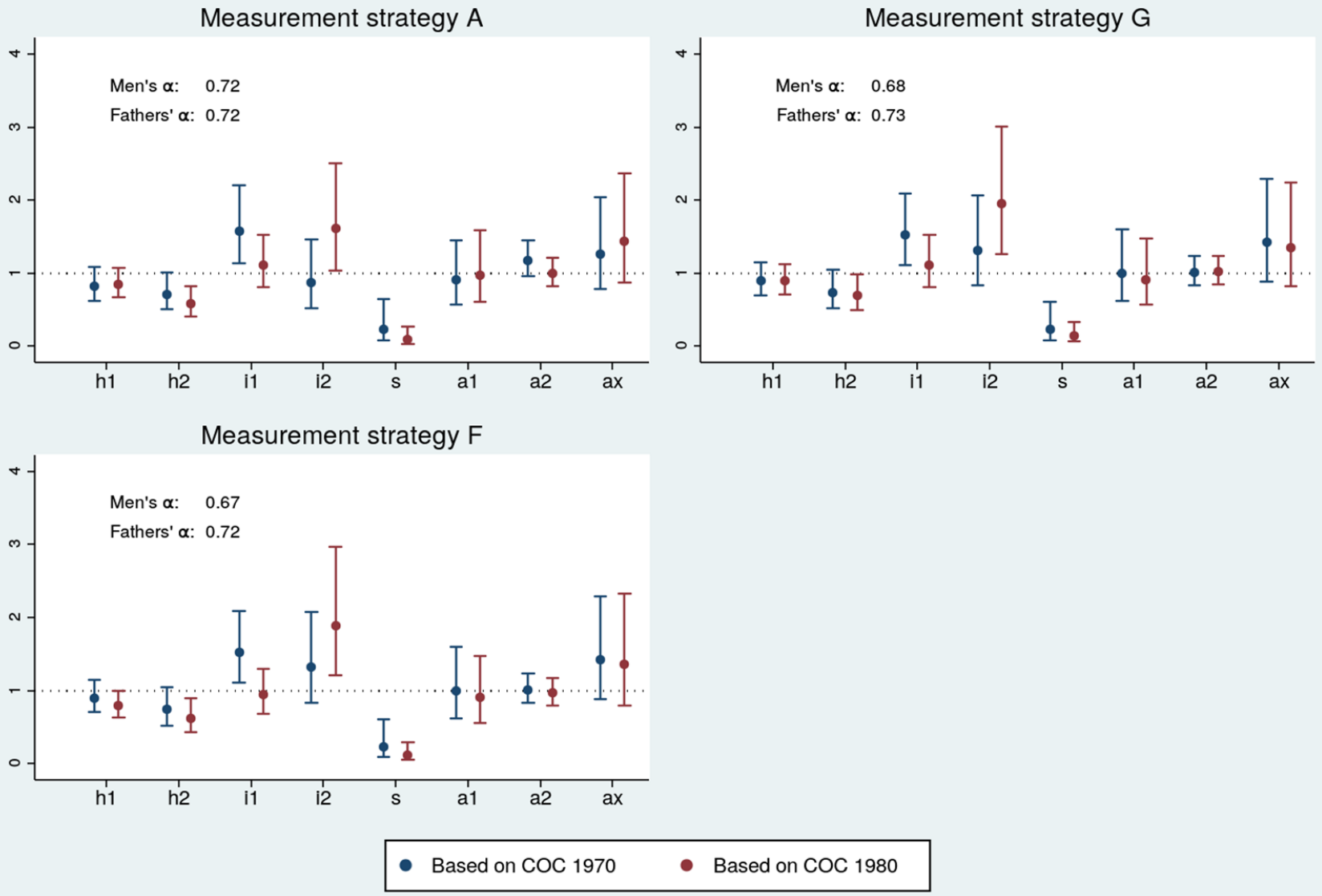


Figure 6: Core-model prediction losses as a function of Krippendorff's α