

Limits of Measurement in the Social Sciences

This conference session on Measurement and Value in Economics presents an excellent opportunity to consider the place of worldview commitments in the work of economists in general and Christian economists in particular. I'd like to organize some thoughts around the three traditional categories of philosophical study: metaphysics, ethics, and epistemology.

Metaphysical discussions

There is a great deal of careful work by Christian economists devoted to considering how the Christian tradition compares to the metaphysical tendencies of modern economics. The modern mainstream of the profession is dominated by the rational-choice framework of neoclassical economics. This framework seems to be built on a particular vision of the human person. Persons are taken to be individuals rather than members of communities; they are self-interested and driven by individual utility-maximization; their preferences are beyond analysis and presumed generally stable; they are moved primarily by financial incentives; their behavior can be studied positively without any appeal to the analyst's norms or values.

All of these statements are metaphysical claims, and to some they would seem to directly contradict much of the Christian tradition. Community obligations, self-sacrifice, conversion to different "preferences," and the centrality of norms and belief (as distinct from "preferences") to the development of real humanity would all seem to be demoted if not challenged by the rational-choice framework. Some economists argue that the dissonance requires Christian economists to forsake the mainstream in favor of a school of thought that is built on different metaphysical foundations. Communitarians have generally favored some flavor of institutional-economics alternative; others advocate versions of Austrian economics.

Many others, however, are content to work as chastened neoclassical economists. Economists model the person as a *utility*-maximizer, not a wealth- or self-maximizer. Utility is a very flexible concept, and it does not necessarily exclude such norms as community obligation or sacrifice. And the rational choice framework is not usually proposed as a *description* of the human person; it is an abstraction, an approximation, with the potential to be useful in analyzing some circumstances and inappropriate for others. For such economists (and I am one of them), the issue is not *whether* to rely on the neoclassical mainstream's approach, but *when*. Some questions (like forecasting demand conditions or evaluating effects of tariffs) seem to prove the mainstream's strengths; other questions (like studying how various welfare policies affect participants' work ethic, or considering the effects of the class structure on economic outcomes) seem to show the limits of the rational-choice framework.

Ethical discussions

There is also a well-developed literature on ethical issues in economics. Some within the mainstream of the profession argue from its metaphysical foundations to conclude that the good society would seem to be one that promotes maximal personal liberty, mobility, and wealth of the autonomous individual, except for cases in which externalities, information scarcity or transaction costs require more cooperation. While some Christian economists find this attractive, much of Christian social teaching has taken care to present a different vision of the common good—one affirming the responsibilities of communities as well as individuals, in which government is viewed as an element of the good creation rather than a result of the fall. This would imply that community responsibilities should not be considered purely voluntary and philanthropic. Though one could not say that a consensus exists, there is a well-developed conversation on the topic.

Epistemological discussions

Say you've navigated the metaphysical discussions--maybe you're a chastened neoclassical economist like me--and you're doing work that is ethically admirable--studying third-world poverty, perhaps, or equality in education. Do metaphysics and ethics exhaust the ways in which the Christian tradition could inform your work? Perhaps not.

Economists are constantly making truth claims--doing epistemology. Usually we're doing this via econometrics. Quantitative studies have become the backbone of modern economic analysis. Our work follows a fairly predictable pattern: Model the most influential variables symbolically, constructing a mathematical representation of the situation under study. Then use inferential statistics to measure the size of the coefficients in the mathematical model. (This requires that we have measured the world reliably and appropriately.) The statistical work is often thought of as a purely technical exercise: Apply the correct technique, and you get truth. Measurement and modeling appear to be value-free; they are techniques that generate their own standards by which to be judged (like t-tests or F-statistics). That seems to settle it: Being a Christian economist means "doing good economics;" the generic values of honesty and technical competence must be honored and applied to worthwhile topics, and that's that.

Christians live in a tradition that has thought long and hard about epistemology, about sound and unsound bases for making truth claims. My basic point in this paper is that I think we could use some more work that considers how we might draw on this Christian tradition to consider how economists support truth claims--to understand the uses and limits of our econometric work, the strengths and weaknesses in the standards to which we usually appeal when trying to establish a given finding. This seems like a potential way to "think Christianly" about things like measurement and

econometric analysis--things that might otherwise seem like technical exercises about which the Christian tradition has nothing much to say.

In this paper I will attempt a very modest contribution to this agenda. First, I'd like to talk through the idea of measurement—some basic measurement theory. Somehow my top-fifteen graduate program in econometrics never touched on measurement theory, and the more of it I read the more I feel chastened. Then I'd like to consider a few cases that illustrate how a little measurement theory might lead us to doubt some of the epistemological habits we may have developed, In the process I hope to also illustrate my belief that empirical work is never purely a technical exercise; values and beliefs always influence even our measurement, which implies that it would be wise to explore how the Christian tradition might inform us about the uses and limits of empirical work.

A Little Measurement Theoryⁱ

Let me start with a few definitions—some names for the types of scales that we typically use to measure things. In general, the measurement scales we most frequently encounter go by five names:

- a. Nominal (or categorical): simple categories for grouping things judged to be in some sense equivalent. These need not even involve numbers. Examples: gender, telephone numbers, type of automobile.
- b. Ordinal: provides a ranking among things that are judged to be in some sense different. (The items being ranked could of course be clusters of things that are all equivalent—an “equivalence class.”) Examples: rankings of economics Ph.D. programs, student letter grades, some measures of air quality. Can't tell us *how much* more of some attribute one item has than the next.
- c. Interval: Like ordinal scales, except that addition and subtraction (but not multiplication or division) of the measures are meaningful. Examples: temperature measured in Fahrenheit or Celsius, calendar time, cardinal utility. Thus “after the cold front came through, the temperature dropped 10 degrees” makes sense, whether the initial temperature was 80 or 40. But it is not true that 80 degrees is twice as hot as 40, because 0 degrees does not measure the absence of heat.
- d. Ratio: Like interval, but multiplication and division are possible because a true zero has been established. Examples: temperature measured in Kelvin, mass, length, time intervals. Non-zero measures here are not “absolute;” 120 seconds is the same as two minutes, so many different scales could be used to measure the same attribute, depending on what we select as one unit of the quality being measured.
- e. Absolute: Measures that don't allow for changes in scale because they are not measured relative to a unit. Examples: number of people in a room, probabilities when taken as relative proportions of outcomes in repeated events.

What is measurement?

At first, “measurement” might have seemed to be one of those self-evident, indefinable ideas. Measurement just means “measuring things.” But we have seen that there are different measurement scales, and each one has different prerequisites; there are underlying circumstances that affect our ability to measure. In the majority of econometric work, including all elasticity estimates, we are presuming we’ve measured items using a ratio scale, so it is worth reflecting upon the conditions that would justify this presumption. Four requirements must be met:

1. We need a precise, unambiguous definition of the characteristic being measured.
2. We need an interpersonal consensus on ranking, relative to the characteristic.
3. We need a set of theoretical properties of the measurement of the characteristic on which a broad consensus is attainable.
4. We need a technique of measurement that is consistent with the definition, the ranking, and the theoretical properties.

To see why these four “prerequisites to measurement” are important, consider a simple case. Say that an economist wanted to study the effect of welfare reform upon the work ethic of the labor force. This economist must now formulate a way to measure work ethic—presumably for individuals as well as for groups. She will have to assert equivalence classes among the persons being studied—collections of states in which an equal amount of “work ethic” is present. Then she will have to determine which states among these equivalence classes present less or more “work ethic” than other states. (Both of those steps concern requirement number two.) All of this will be impossible without a clear definition of “work ethic” (requirement number one). But if this exercise is to be at all useful to any community, the measurement must not be merely based on her own private definitions and convictions; there must be a general agreement on the meaning of “work ethic,” *and also* on the various amounts of this work ethic that are present in various states of the persons being studied. (That’s still requirement number two.) Without a broad agreement on this topic, there is no use trying to move beyond a purely subjective assertion about individuals’ work ethic; no measurement, not even an ordinal ranking, is possible. (And we haven’t even gotten to requirements three and four!)

That is a great deal to presume about the notion of work ethic, but even this level of agreement only allows an *ordinal ranking* of individuals’ work ethics. If we are to move on to an empirical study, measuring elasticities and so forth, we must meet the third and fourth prerequisites listed above. We need a broadly accepted theory of the characteristic we’re calling “work ethic.” So we need to establish a true zero for the characteristic (i.e., what level of “work ethic” is equivalent to having none at all?), and also an interval for measurement (i.e., how much more of “work ethic” must be present before we would say a person has risen from a zero to a one?). Then we will need a well-developed theory that indicates how our measurement of work ethic behaves over the whole range of human experience. (For example, is the person with a measured level of work ethic equal to 0.5 really half-way between the person

measured to be a zero and the person measured to be a one? How about the person with a work-ethic measurement of 5.7? Does this measurement really correspond to 5.7 times the work ethic of the person registering a one on our measurement scale?)

You can see why economists run into difficulties when considering norms and preferences. We can't get past the second of our four prerequisites for measurement when it comes to preferences, so we must be content to rely on only ordinal scales—rankings of preferences. That's why our bright students turn up their noses when we start to talk about "utils" in the first weeks of microeconomic theory—they know there is no unit for measuring preferences or beliefs.

When we do an empirical study that, say, reports elasticities reflecting the effect of social policy on work ethic, we are presuming we can go well beyond such ordinal scales, to measure work ethic on a ratio scale—a scale that includes a meaningful zero and a meaningful unit. We're also assuming that there is a well-accepted theory that links the measurement to the measured characteristic in a linear way. When we report price elasticities, we are presuming the same kind of thing regarding scarcity—we treat prices as ratio-scale measurements of relative scarcity, which presumes we have connected the idea of scarcity to the measurement of price in a way that satisfies our four measurement prerequisites. When we measure consumer and producer surplus as an index of the relative gains from market exchange, we make multiple strong claims—not only must price meet high standards as a measure of scarcity, but our demand- and supply-curve coefficients must meet the same standards as measures of the way that quantities *would* change if price and other influences were to change.

These are heroic presumptions, well worth some reflection. But they are only the beginning, because measurement is part of a larger process, the process of seeking information that we might call "the information cycle:"

- Before studying anything, we must identify a need for information.
- Then a precise question must be formulated.
- We then select a measurement technique—the process I've been describing.
- We collect the data we believe to be relevant. We engage in some exploratory data analysis, beginning to formulate hypotheses and models.
- We test hypotheses and begin to generate some results.
- Finally decisions are made and policies are constructed on the basis of our results.

Each of these steps in the information cycle involves decisions that are influenced by the norms and beliefs of the analyst. And there are feedbacks to be considered among the steps. For example, often collecting the data we'd like to have proves to be impossible, sending us back to the previous step in the cycle, or forcing us to casually choose a proxy measurement that may meet none of the standards we really should be seeking. Economists do this all the time. Work ethic is proxied by labor participation rate; well-being is proxied by GDP; poverty, a complex social and cultural phenomenon, is proxied by relative income level. In each case, the proxy is chosen not because it meets the normal prerequisites

for measurement, but because it seems to be “measurable.” It’s handy. But great is the potential for drawing bad conclusions through the use of a bad proxy or mis-measured attribute.

A Case Or Two

By now we have several issues to think through: the ways in which values enter the various steps of the information cycle, the ways in which economists sometimes use variables that do not meet the necessary prerequisites for measuring the attributes under study, and the ways in which inappropriate proxies are often chosen out of convenience. Consider three brief cases in which some thought about epistemology would improve our work, and in which technique does not generate the standards for its own evaluation. I realize that some of these cases might be dismissed as not representative of the broad issues that economists face, but I believe that is not the case. I also appreciate that my comments might have the ring of common sense that any practicing economist already knows, but in my experience this is also not the case.

What Is a Statistical Test?

Statistical tests, especially the t-statistic, are probably the single most frequently-cited basis for economists’ epistemological claims. How do we know X, Y or Z? Usually because a t-statistic told us so. We’ve done a regression, and the coefficient on a particular parameter was “significant,” according to its t-statistic.

In the standard textbook presentation of statistical methods, based on the presumption of scientifically controlled and repeatable experiments, a person hopes to actually measure uncertainty and narrow it by sampling repeatedly from a stable population. In this classical case, probabilities are defined as the relative frequencies with which events occur after repeated sampling.

Because social scientists usually can not repeat events under the same conditions, they are modeling the “unsameness” of events by including the possibility of random white noise. This is justified by the belief that we have organized the variables (i.e., specified the model) so that this noise (unspecified influences and approximations) has, on average, no influence on the relationships among the variables we are studying.

Now exactly what does “random” mean when applied to the non-experimental data of this example?ⁱⁱ The data are not the result of a repeatable random draw from a stable population. Remember, we are not really involved in experiments. From what “population” were the data drawn for this month’s number of homicides? Here the “sample” is the population, and the current situation will not be repeated in a future month, from which another random sample might be drawn. In fact, by the time the data are generated the events in question have already occurred. So in what sense can we talk about their “relative frequency”? An event does not have much frequency if we can be certain that it will only happen once, in the past.

Therefore our actual practice implies that we are giving non-classical meanings to the words "probabilities" and "randomness," due to our dependence upon non-experimental data. Statistical methods are being used not only to summarize data, but also to simulate the control groups and random sampling that are essential to scientific work. But this has implications for the way we should define "probability" and "random," as well as the way we should interpret hypothesis tests and confidence intervals. These tests and intervals also take on different meanings than in classical statistics.

It is clear that we can not mean the word "random" in a literal sense here. "Random error" is being used as a metaphor; it is a way to model our ignorance. "Random" does not (as in classical statistics) refer only to observations and events; it is actually our knowledge of events, our model and measurement of them, that is random, influenced by errors and ignorance. We are modeling all of this as randomness. Our choices of which variables to exclude and which functional forms to use and how to measure proxies are all made subject to error, and these errors will create noise that knocks our observations away from the places a good theory would predict they should be. The best we can hope for is that we will make these errors in a way that leaves the resulting noise non-systematic--that is, not polluting our estimates of the coefficients in the model.

It is all right, then, to speak metaphorically of the "chance" that a non-repeatable social science event would be observed, or the "probability" that a parameter takes a particular value, *if* we are careful about the meanings of chance and probability. Classical statistics, using experimental data, yields the likelihood that we have correctly accepted or rejected a null hypothesis, but with non-experimental data the null hypothesis itself has a "probability" or degree-of-belief attached to it. Probabilities here are degrees of warranted belief, not relative frequencies. Probabilities are indexes of the reasonableness-of-doubt that should be attached to our conclusions--conclusions which, as in court cases, can not generally be proved or disproved, but can be argued more or less persuasively. Probability is not a statement about physical events, but an estimate of the level of believability, the relative weight of admissible evidence in the face of uncertainty and ignorance; it is an assessment of the likelihood of a particular conclusion, given a body of (imperfect) information.

Probabilities in the social sciences are therefore formal ways of translating degrees of reasonable belief into real numbers. These probability numbers then conform to several norms of measurement regarding a proper "probability measure;" for example, probabilities of all possible outcomes must sum to one.

This all amounts to a method for moving from an ordinal ranking to a ratio scale measurement of subjective degrees of belief. But—and here is my point!--this raises a serious problem for the use of classical probability theory in the social sciences: We have argued that ratio scales are only appropriate if one has a well-worked-out theory of the entity being measured, a "correspondence theorem" that justifies the claim that the entity in fact behaves like the real numbers. Do we have such a theory of "subjective degrees of belief?" Can we, for example, be confident that a result that bears a probability of .42 is actually "half as believable" as a result that is .84 probable? And is this confidence

intersubjective-- true for all reasonable observers of the events? I am not aware of any satisfactory correspondence theorem for subjective degrees of belief,ⁱⁱⁱ even though all of our usual use of statistical tests would require such a theorem.

This does not imply that formal, empirical methods are not appropriate in the social sciences. The question becomes not one of identifying types of problems for which formal methods are not appropriate, and others for which they will yield the right answers. By themselves the methods can never be relied upon to yield correct answers. Like the act of collecting footprints or bullet fragments, formal methods can clarify reality, or obscure it, or create an alternative virtual reality that misleads. It is still common sense, creativity and discernment that must be invoked when a practitioner sits down to a problem. Formal methods can offer contributory evidence, but not conclusive demonstrations.

Instead of asking when (or if) formal methods yield the right answer, we should ask when to put a high degree of confidence in the limited information we get from formal methods. When one has solid data that conform to a measurement scale appropriate to the topic, clear contestable predictions, few auxiliary hypotheses, and a well-behaved error term, one can afford to be less skeptical. Before invoking statistical tests of hypotheses, we would do well to ask if we know the situation being modeled, and if fair persons with similar knowledge would trust the general approach to the problem and the results obtained. Levels of statistical significance are always somewhat arbitrary, but we should be especially skeptical in cases (1) where the social processes under study are extremely complex, with many auxiliary hypotheses complicating the primary hypotheses; (2) where the entity being measured is not clearly definable, or there is a poorly developed theory of the entity and its relationship to the measurable variables, or the measurement instrument is not precise and reliable; (3) where inappropriate measurement scales are used; (4) where the statistical methods (and, where present, functional forms) employed are not consistent with the measurement scale; (5) where the specification of the model and its functional form are not clearly justified by reference to the actual situation being modeled; (6) where the error residuals are not observed and analyzed, as in some ANOVA and correlation studies; and (7) where the quality of the data and reliability of the source are questionable. We should be particularly skeptical when the analyst does not fully disclose the relevant information on these topics. In fact, it should be a professional norm that the statement of one's results must, as a matter of habit, discuss these issues.

Welfare Reform

Welfare reform is perhaps *the* economic story of the last decade. We have experienced a sea change in social economic policies, and the change has potentially huge effects upon the most vulnerable in society.

How should we evaluate welfare reform? Consider for a moment why we have welfare programs in the

first place. We believe that to some extent hard times come on people randomly, through disability or economic changes, so it makes sense to share these risks. In other cases, poverty is not random, but systematic to the point of seeming deliberate, and we want to fight the effects of racism, gender discrimination, and unfair privilege. We believe that in other cases children are innocently affected by the unwise choices of their parents, and want to mitigate this harm as a matter of fairness.

So to evaluate welfare reform, what things should we want to measure? Changes in the incidence of economic vulnerability; changes in the stability of family income, or at least its level; changes in the incidence of sexual or racial discrimination; changes in the wellbeing of children; changes in the incentives toward or incidence of unwise parental choices.

Now consider some quotes from an excellent welfare-reform literature review article recently commissioned by the *Journal of Economic Literature*.^{iv}

1. "The most voluminous literature on welfare reform in the past decade has focused on caseload changes." (Blank, 37)
2. "It is important to explore not just whether policies caused women to leave welfare, but also whether these same policies helped women enter the labor force. ... Labor market opportunities for this group have received substantial public attention." (Blank, 51)
3. "The net income-increasing or poverty-reducing impacts... (are) a much less-well understood area of analysis than the analysis of caseload change or labor force participation." (Blank, 57) "In part, this reflects the fundamental problem of appropriately measuring economic well-being. ... I summarize what is known about the interaction between welfare reform and changes in income, poverty, and other measures of well-being. The existing information described in this section is quite limited." (Blank, 58)
4. "Legal immigrants entering the country after August 1996 were made ineligible for virtually all forms of federal public assistance... There has been remarkably little literature studying the impact of these provisions on the behavior and well-being of the immigrant population." (Blank, 62)
5. "Evidence is limited that relates welfare reform more broadly with child outcomes... This is another area where further research would be highly useful." (Blank, 63)

In other words, faced with a fundamental change in social policy, we economists come to the profound conclusion that, if you make people ineligible for welfare, they will use it less and work more. We really aren't sure about everything else. Why? In large part it's an issue of measurement. Our work could be improved by attending to the epistemological limits of the tools to which we are accustomed. I'm not sure we even think about why we are measuring certain things or relying on particular proxies rather than pressing for measurement of the actual attribute we wish to study. And if there were no proxy problem we would still need to wonder if we're measuring the items of interest properly. In the present

case, for example, we crank out welfare-related labor supply elasticities in great number, even though Moffitt's classic 1983 *AER* article demonstrates that the presence of welfare stigma leaves them all biased.^v Mark Twain's comment about the weather applies to Moffitt's article as well: Everybody talks about it, but nobody *does* anything about it.

Discrimination Lawsuits

Let's finish with a short empirical example. SPSS Version 10.0 provides a dataset of employee-related data from a single firm (N=474), reporting the following variables:

GENDER	Male/female employee gender, coded male=1
EDUC	Educational level, years
JOBCAT	Employment category: 1 = Clerical 2 = Custodial 3 = Managerial
SALARY	Current salary, dollars
SALBEGIN	Beginning salary, dollars
JOBTIME	Months since hire
PREVEXP	Previous work experience, months
MINORITY	Minority status; coded 1=minority

Say that we pose the following question: Is this firm's compensation regime fair? If we judged that a fair firm would compensate its employees based on their experience, job tenure, education, and the responsibilities of their job classification, we might construct a regression like this:

Table One: Parsimonious Model

Coefficients				
	Coefficients	Std. Error	t	Sig.
(Constant)	-21023.630	3942.120	-5.333	.000
Educational Level (years)	2065.544	173.285	11.920	.000
Months since Hire	109.232	42.626	2.563	.011
Employment Category	13260.687	645.815	20.533	.000

a Dependent Variable: Current Salary

Look at those wonderful t-statistics! Every variable is significant at roughly the 99% confidence level. And the F-test for significance of the regression as a whole is a whopping 373.23, significant at a better-than .999 level. All of the coefficients also have the "right" sign. (We still might wonder about the serious measurement problem raised by our treatment of job category: We've included three categories by

measuring them 1-2-3, as if the third category were known to have 3/2 as much productivity, worker scarcity and value as the second category.)

Unfortunately, these results do not answer our research question. They do not, for example, tell us if people are sorted into job employment categories in a fair, non-discriminatory way. We do not have any direct information about sorting into job categories, but we might get a weak proxy by asking how previous experience is rewarded, relative to employees who rise internally through the old-boy network. If we include the “months of previous experience (before hire)” variable, the F-statistic is changed very little, and the t-statistics remain highly significant, but the coefficients begin to tell a new story:

Table Two: Full Productivity-based Model

Coefficients

	Coefficients	Std. Error	t	Sig.
(Constant)	-19119.788	4042.404	-4.730	.000
Educational Level (years)	1943.168	183.179	10.608	.000
Months since Hire	111.037	42.499	2.613	.009
Employment Category	13568.931	661.820	20.502	.000
Previous Experience (months)	-8.703	4.337	-2.007	.045

a Dependent Variable: Current Salary

Experience outside of this firm is being penalized; it bears a negative coefficient.

Now if we add consideration of race and gender, we again get an F-statistic that is significant at the .999% level and t-statistics that are significant near the .99% level, but we draw a very different conclusion about the fairness of the firm:

Table Three: Full Productivity-based Model Plus Discrimination Variables

Coefficients

	Coefficients		t	Sig.
	B	Std. Error		
(Constant)	-15461.374	3983.238	-3.882	.000
Educational Level (years)	1671.481	185.104	9.030	.000
Months since Hire	103.850	41.390	2.509	.012
Employment Category	12724.601	661.698	19.230	.000
Previous Experience (months)	-12.695	4.363	-2.910	.004
Minority Classification	-2461.421	1034.077	-2.380	.018
Gender	4968.513	954.312	5.206	.000

a Dependent Variable: Current Salary

Without much of a change in the first three coefficients relative to the initial regression, we now learn that women are penalized roughly \$5000 per year for their gender, and minorities roughly \$2500 per year for their race.

My point is not that any working labor economist would fail to consider race or gender in an earnings equation. Of course he would. But consider the statistical tests I've just reported. My point is that tests of significance really tell us *relatively little about the significance of our findings*. They do not tell us if the correct variables have been included; they don't tell us if the variables have been measured properly; they don't tell us if we should add or delete variables from our model; they don't tell us if our model is reasonable.

But unfortunately that is just the beginning. If we were really serious about this firm's hiring practices—that is, if our money were on the line—we would probably be in court. And, as it happens, in cases of this type a different criticism of the t-test has arisen. We've approached the firm as if salary were the dependent variable, with non-stochastic explanatory variables like experience and education. It has become not uncommon for one side of discrimination-related lawsuits to view the situation in this way, with the other side invoking the need for a reverse regression: It might (at least in principle) be the case that job openings are posted with a fixed salary offer, after which applicants with various levels of education and experience make application. Thus salary should be considered the non-stochastic independent variable, and the formerly-independent variables should be considered stochastic and dependent.

This creates a logistical problem: How can we construct a single dependent variable from the several formerly-independent variables, in order to complete the reverse regression? Typically a log-linear regression of salary on productivity measures (performance ratings, years of experience, full-time/part-time status) and discrimination-related binary variables (for race and/or gender) is estimated. Then an index of productivity is calculated. Normally, productivity is indexed by the forecast value of $\ln[\text{Salary}]$ from the first regression, using forecasts that ignore the contribution of race to salary. A reverse regression of the productivity index on log salary is then completed.

One of the fascinating themes in this branch of the forensic-economics literature is that the two approaches, when given identical data, generate tests of significance that very frequently draw *opposite conclusions* about the presence of discrimination. In their recent review article which also included original regressions for a particular firm, White and Piette find that direct regression leads to a conclusion that discrimination exists within the department, while the reverse-regression analysis leads to "results (that) are completely opposite from those of the direct regression approach."^{vi} They argue that this paradox is not an artifact of their particular data set, finding that "it is common to find the implications of the direct regression results opposite of those from the reverse regression."^{vii}

The following two tables present log-linear regressions using the SPSS data--first for a direct regression, then for a reverse regression using the same data. The direct regression indicates that males and non-minorities have a salary advantage after all other explanatory variables have been

accounted for. These discrimination-related coefficients, like all of the others in the regression, are significant at a 99% level of confidence. The reverse regression indicates that males have lower qualifications at any given salary level than females, but this coefficient fails to achieve the level of confidence of the direct regression. The reverse regression yields a positive coefficient on minority status, indicating that non-majority workers have higher qualifications at any given salary level, though this coefficient bears a terrible t-test for significance (P-value = .978). In the literature, results like those in these tables are routinely interpreted as providing support for both parties in salary-fairness litigation.

Table Four: Direct Regression Results, Log-linear Model

Source	SS	df	MS	Number of obs = 474		
Model	57.6151857	6	9.60253096	F(6, 467)	=	262.87
Residual	17.0594343	467	.036529838	Prob > F	=	0.0000
				R-squared	=	0.7715
				Adj R-squared	=	0.7686
Total	74.67462	473	.157874461	Root MSE	=	.19113

lsalary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.0420381	.0039229	10.72	0.000	.0343293	.0497469
prevexp	-.0003981	.0000925	-4.31	0.000	-.0005798	-.0002164
jobcat	.2704028	.0140235	19.28	0.000	.2428458	.2979597
jobtime	.0025214	.0008772	2.87	0.004	.0007976	.0042451
gender	.1822294	.0202249	9.01	0.000	.1424863	.2219725
minority	-.0741008	.0219154	-3.38	0.001	-.1171657	-.0310359
_cons	9.158715	.0844174	108.49	0.000	8.99283	9.3246

Table Five: Reverse Regression Results, Log-linear Model

Source	SS	df	MS	Number of obs = 474		
Model	30.6005274	3	10.2001758	F(3, 470)	=	423.02
Residual	11.3331213	470	.024113024	Prob > F	=	0.0000
				R-squared	=	0.7297
				Adj R-squared	=	0.7280
Total	41.9336487	473	.088654648	Root MSE	=	.15528

qualif	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lsalary	.6643316	.0217821	30.50	0.000	.6215293	.7071338
gender	-.0392546	.0170984	-2.30	0.022	-.0728534	-.0056559
minority	.0004977	.0179318	0.03	0.978	-.0347387	.0357341
_cons	3.414776	.221865	15.39	0.000	2.978806	3.850746

This would all seem to make reverse regression a litigant's dream: an established statistical technique that will produce defensible evidence for either side of a discrimination lawsuit. A litigant's dream, and a philosopher's nightmare. One hopes that there really is an objective truth "out there" to be explored via statistical work. Yet a review of the literature might leave the impression that, in

discrimination lawsuits at least, mathematics is not much more than a rhetorical device in the service of advocacy.

We might avoid such cynicism by a little attention to the epistemological limits of our methods--especially our tests of significance. Maddala's^{viii} discussion of reverse regression indicates that "since the direct regression gives biased estimates of (the qualification coefficients), what we have here is a biased index of qualifications," so that "one should not make inferences... but obtain bounds for (the discrimination coefficient) from the direct regression and reverse regression estimates." In other words, since the dependent variable in the reverse regression is subject to estimation errors, statistical tests in the reverse regression are not valid. In the present example, the statistical insignificance of the reverse-regression's race and gender coefficients are irrelevant to the issue under study. The t-test just isn't appropriate in such a regression. Maddala^{ix} (1992, 459-461, 71-74) concedes that there are cases in which both direct and reverse regressions must be employed, since in these cases one approach may tend to overestimate the crucial coefficients while the other may tend to underestimate them. In such cases we should think of the discrimination-related coefficients of the two regressions as upper- and lower-bound measures of the true extent of discrimination. In either event, the two regressions I've presented actually give unambiguous evidence of the presence of discrimination.

Conclusion

Econometric methods have allowed for some genuine breakthroughs in analysis, resulting in real improvements in the world. I am not making an argument that they should be abandoned by economists who seek to be faithful to their Christian convictions. But any good tool can be used unwisely, and I have suggested that we will be less susceptible to the unwise use of statistical methods if we think of them as a form of epistemology. This brings our use of econometrics and measurement into a long conversation about how truth claims should be evaluated. I think that, if we pursue this line of thinking, we may find that the Christian tradition will yield insights for wisdom in the use of econometrics and economic measurement, generating an expanded sense of the value and limitations of the ways in which economists usually do their work.

ⁱ For a fine review of measurement theory, see Henry Kyberg, *Theory and Measurement*, Cambridge, UK: Cambridge University Press, 1984.

ⁱⁱ This section is in debt to the detailed discussion of the notion of "randomness" in non-experimental settings in Chapter One of A.C. Darnell and J.L. Evans, *The Limits of Econometrics*, Brookfield, VT: Edward Elgar, 1990, Chapter One.

ⁱⁱⁱ One could also make the argument that measurement theory indicates that ratio scales, such as probabilities, are strictly inappropriate for non-experimental data where probabilities index degrees of reasonable belief.

^{iv} Blank, Rebecca M. "Evaluating Welfare Reform in the United States." NBER Working Paper Series, copyrighted 2002.

^v Moffitt, Robert. "An Economic Model of Welfare Stigma." *The American Economic Review*, 73/5, December 1983, 1023-1035.

^{vi} P.F. White and M.J. Piette, "The Use of 'Reverse Regression' in Employment Discrimination Analysis," *Journal of Forensic Economics*, 11(2), 1998, 127-138, p.133

^{vii} *ibid.*

^{viii} G.S. Maddala, *Introduction to Econometrics*. New York: Macmillan, 1992, 461.

^{ix} *ibid.*, 459-461, 71-74.