Philosophy of Science (2025), **92**, 285–303
doi:10.1017/psa.2024.25

PSA PHILOSOPHY OF SCIENCE ASSOCIATION
PHILOSOPHY OF SCIENCE

ARTICLE

The Logit Model Measurement Problem

Stella Fillmore-Patrick

School of Historical, Philosophical, and Religious Studies, Arizona State University, Tempe, AZ, USA
Email: stella.fillmore-patrick@asu.edu

(Received 24 October 2023; revised 11 April 2024; accepted 31 May 2024; first published online 05 December 2024)

Abstract

Traditional wisdom dictates that statistical model outputs are estimates, not measurements. Despite this, statistical models are employed as measurement instruments in the social sciences. In this article, I scrutinize the use of a specific model—the logit model—for psychological measurement. Given the adoption of a criterion for measurement that I call comparability, I show that the logit model fails to yield measurements due to properties that follow from its fixed residual variance.

1. Introduction

Do statistical models yield measurements? On the face of it, clearly not. Consider an example: suppose I need to measure the square footage of my apartment. To do this, I decide to construct a statistical model predicting square footage given the number of bathrooms. I gather an immense amount of data on the number of bathrooms and the square footage of apartments. I have one bathroom, and my model estimates that my apartment is 621 square feet. It seems wrong to claim to have measured the area of my apartment. Say that, as it turns out, my apartment is in fact 621 square feet. Even then, it seems wrong to say that I have measured the square footage of my apartment. Statistical model outputs are usually described as estimates or predictions, not as measurements.

Surprisingly, some do treat some statistical models as measurement instruments. Social scientists like psychologists and economists study attributes that are not observable: intelligence, well-being, market value, inflation, and so on. Lacking the option of direct physical measurement, researchers create models to represent and measure these hidden magnitudes. Using statistical models as measurement instruments is a pragmatic response to the circumstances in these fields.

Despite this practical necessity, treating statistical models as measurement instruments is counterintuitive. And so, in this article, I take up the issue of whether statistical models are measurement instruments. Because statistical models are varied in their properties, rather than considering statistical models in general, I consider whether one *particular* statistical model that is important for psychological

measurement can be treated as a measurement: the logit model. Ultimately, I argue that the logit model is not a measurement instrument because it violates a necessary condition for measurement: comparability of measurement outcomes.

Here is what I plan to do in this article: in section 2, I give some necessary technical background. In section 3, I show that two different interpretations of the logit model developed out of an old debate between George Udny Yule and Karl Pearson about how to measure association between binary variables. I trace the use of the logit model as a measurement instrument to one of these interpretations, though I will ultimately argue that this use is misguided. In section 4, I argue that measurement instruments necessarily produce comparable outcomes. I call this the *comparability* criterion. I show that a variety of philosophical views on measurement can adopt comparability without challenge to their position. Finally, I argue that the logit model violates the comparability criterion and thus cannot be used for measurement.

2. Background

Statistical models are used to measure unobservable traits in psychology. Because these traits are unobservable, the models that are used to measure them are called *latent variable models*. A latent variable is one that is unobserved, meaning that no data correspond to the variable. One such latent variable model that is important for psychological measurement is logistic regression, also known as the logit model. In this section, I explain what logistic regression is and how it can be understood as a representation of an unobserved variable that is supposed to exist.

2.1. Linear regression

Regression analysis uses an intuitive idea to discover relationships in indeterminate data. Simply put, a linear regression takes a scatterplot and finds a line of best fit. In this way, indeterminate relationships are discovered between features represented on the x and y axes. Imagine, for example, that a scatterplot is constructed to plot a man's height (y) against the height of his father (x).[1] The relationship between x and y is not determinate, meaning that given the height of a man's father, one cannot determine the height of the man. Knowing the father's height, though, will allow one to make a better estimate of the son's.

In this example of a simple linear regression, the father's height is a *predictor variable*, while the son's height is an *outcome variable*. The mathematical relationship between the predictor variables and the outcome variable (this is typically the variable that most interests the practitioner) is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_j X_{ij} + \epsilon_i, \tag{1}$$

where Y_i is the outcome variable, X_{ij} is the predictor variable, β_j is a coefficient, and ϵ_i is the error or difference between the predicted value and the actual value. Given

[1] Francis Galton used linear regression to model the relationship between the heights of fathers and sons. He found that very tall fathers would have sons with heights closer to the average, and ditto for very short fathers. This is why linear regression is called "regression": Galton saw that there is a regression to the mean. This also explains the "sophomore slump" in sports.

https://doi.org/10.1017/psa.2024.25 Published online by Cambridge University Press

some data, the coefficients are estimated to minimize the difference between the model predictions and the observed values for the outcome variable. The value of the coefficient characterizes the relationship between the outcome variable and the predictor variable. For example, if a regression model predicts a man's height given the height of his father, the coefficient represents the change in a man's height per unit increase in the height of his father.

However, linear regression cannot model binary outcome variables like sex or mortality because it is a continuous linear function. Logistic regression was developed to address this limitation.

2.2. Logistic regression

The general strategy of logistic regression is to transform the data such that the outcome variable is the log of the odds of a particular outcome rather than a predicted value. For example, if mortality is the outcome variable of a logistic regression, the model yields the log of "odds of mortality" given the predictor variables rather than mortality directly. This is convenient because odds are continuous and so can be modeled with a continuous linear function.

In a logistic regression, the log of the odds (called a *logit*) is predicted rather than the odds. Whereas odds can range between 0 and ∞, the log of odds ranges from $-\infty$ to ∞, so the regression model will always produce an outcome that is within range.

The logistic regression applies the following transformation to the outcome variable:

$$\text{logit}(x) = \log\left(\frac{x}{1-x}\right). \tag{2}$$

In a logistic regression, it is stipulated that the regression equation estimates the transformed outcome variable rather than the outcome variable directly. The regression equation is then rewritten as

$$\text{logit}(P(Y_i = 1)) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_j X_{ij}, \tag{3}$$

where Y_i is the binary outcome variable, X_{ij} is the predictor variable, and β_j is a coefficient. The coefficients again represent the relationship between the outcome variable and the predictor variable, although in logistic regression, they are interpreted as a change in log-odds of success given a one-unit change in the predictor variable. Unlike the linear regression, there is no error term in the mathematical form of the logit model. The indeterminacy of the model is instead captured in the probabilistic interpretation of the outcome variable.

To summarize, the logit model applies the regression framework to binary outcome variables by estimating the log-odds of an outcome rather than estimating the outcome directly.

3. Early interpretative disputes about the logit model

In what precedes, I gave a brief overview of the logit model. In this section, I explain how the logit model came to be viewed as a measurement instrument by social scientists like psychometricians, those who measure psychological attributes.

https://doi.org/10.1017/psa.2024.25 Published online by Cambridge University Press

Table 1. Contingency table example with synthetic data

	Philosophy PhD = 1	Philosophy PhD = 0
External world belief = 1	15	970
External world belief = 0	5	10

A historical dispute about measuring association among binary variables can provide insight into this development.

In the twentieth century, statisticians George Udny Yule and Karl Pearson proposed different ways to measure the degree of association between binary variables. Yule, a former student of Pearson's, proposed to measure association with relative odds, while Pearson developed a measure that depends on positing an underlying (latent) continuous variable to explain the observed binary data. The differences in the measures of association revealed deep philosophical disagreements between Yule and Pearson that I discuss in this section.

As Powers and Xie, (2008) have observed, a parallel dichotomy emerged from the Yule/Pearson debate for the logit model. According to one interpretation of the logit model, the outcome is just a log-odds. This interpretation, called the *transformational approach*, resembles Yule's approach to measuring association. According to another, the outcome variable represents an unobserved variable that explains the behavior of the manifest binary variable. This interpretation, called the *latent variable approach*, resembles Pearson's approach to measuring association. The second interpretation, the one that appeals to an unobserved variable, provides an initially plausible (but ultimately inadequate) justification for the use of the logit model for measurement in the social sciences.

3.1. Yule's measure of association

George Udny Yule, (1900) developed his measure of association for binary variables based on change in odds. This measure is called *Yule's Q*. Like correlation, it is a measure between −1 and 1 of the relationship between two variables. A measure of 0 indicates that the variables are independent, whereas a measure of −1 or 1 indicates that one can infer the value of one variable from the value of the other variable. The odds ratio (OR) scaled between −1 and 1 is Yule's Q:

$$Q = \frac{OR - 1}{OR + 1}. \tag{4}$$

Suppose we have a data set of 1,000 observations with two binary features: whether or not one has a PhD in philosophy (philosophy PhD) and whether or not one believes in the external world (external world belief).[2] The data are organized in a table with four cells, such that each combination of values is accounted for (see table 1).

[2] These data are made up.

https://doi.org/10.1017/psa.2024.25 Published online by Cambridge University Press

To calculate the measure of association between philosophy PhD and external world belief using Yule's Q, one must measure whether one is more or less likely to believe in the external world given that one has a PhD in philosophy compared to not having a PhD in philosophy. This is calculated and standardized as follows:

$$Q = \frac{(15)(10) - (970)(5)}{(15)(10) + (970)(5)} = -0.94, \tag{5}$$

where $Q = -0.94$ indicates that the binary variables are strongly negatively associated, meaning that if one has a philosophy PhD, the odds of not believing in the external world increase relative to the odds if one doesn't have a philosophy PhD.

Yule's Q is based on the OR. The OR is a measure of how the odds of one variable change relative to the value of a second variable. For example, the OR for our toy example is

$$\text{OR} = \frac{\big(15/(15 + 970)\big)/\big(970/(15 + 970)\big)}{\big(5/(5 + 10)\big)/\big(10/(5 + 10)\big)}. \tag{6}$$

The numerator is the odds that 'Philosophy PhD' = 1 given that 'External world' = 1. The denominator is the odds that 'Philosophy PhD' = 1 given that 'External world' = 0. The odds ratio can be simplified to the ratio of the cross diagonals:

$$\text{OR} = \frac{(15)(10)}{(970)(5)}. \tag{7}$$

3.2. Pearson's measure of association

Pearson's measure of association between binary variables, on the other hand, is based on the assumption that there are underlying continuous variables that govern the observed binary variables. If, for example, the binary variable observed is mortality, some underlying continuous but unobserved variable (perhaps overall health) is assumed to account for the value of the binary variable mortality. At a certain level of the unobserved variable, the observed variable will change values. On the basis of the observed binary data, an underlying continuous distribution is fit that would give rise to the binary observations. The properties of the underlying distribution of each binary variable are used to construct a measure of association between −1 and 1 Pearson, (1904).

This method depends on distributional assumptions, whereas Yule's Q does not.[3] Despite the fact that his measure introduces more assumptions, Pearson preferred this because it closely mirrors his measure of correlation between continuous variables. Pearson's correlation coefficient for continuous variables assumes a bivariate normal distribution among the variables (for the most meaningful result). It

[3] A distributional assumption mathematically characterizes the data-generating process that gives rise to the sample data. Making a distributional assumption will constrain the extent to which one's model is determined by the observed data, because one's data will contribute to parameter estimates within a particular mathematical form (distribution). For example, if I assume that my data are normal, my model will look like a bell curve no matter what. The data will determine the width and height of the bell curve.

https://doi.org/10.1017/psa.2024.25 Published online by Cambridge University Press

measures the strength of a linear relationship among the data-generating distributions, not merely among the observed data.

Pearson has other reasons for preferring his measure over Yule's. Pearson and Heron (1913) argue that classifications that appear to be binary actually track continuous quantities. In the case of mortality, for example, they claim that binary data recording death or recovery "were used to measure a continuous quantity—the severity of the attack" (192).

3.3. Reflecting on the Yule/Pearson dispute

Yule was concerned that the additional distributional assumptions suggested by Pearson licensed inferences that went far beyond the observed data. He criticized Pearson in a 1912 paper, stating that "apart altogether from the question of its legitimacy, the assumption [of an underlying continuous distribution] is unnecessary and unverifiable" (612). Given that Pearson's underlying distributions could not be verified, Yule suggested that his own measure of association, based purely on the observed data, should be preferred.

Pearson replied, "If Mr. Yule's views are accepted, irreparable damage will be done to the growth of modern statistical theory" (Pearson and Heron 1913, 158). Continuing his response, he states that "the controversy between us is much more important than an idle reader will at once comprehend. It is the old controversy of nominalism against realism. Mr. Yule is juggling with class-names as if they represented real entities, and his statistics are only a form of symbolic logic. No knowledge of a practical kind ever came out of these logical theories" (302).

While Yule is motivated by epistemic conservatism, Pearson argues that classification that does not rest on some underlying continuous variation is meaningless. Thus he accuses Yule of nominalism, while he sees himself as defending scientific realism. Pearson never directly replies to Yule's epistemological point, however. Yule's point is that *even if binary outcomes are in fact grounded in underlying continuous attributes*, the observed discrete data underdetermine them, and their existence cannot be verified.

The issues on display in this old dispute continue to plague the interpretation of statistical techniques in a new context—psychological measurement. The two competing interpretations of the logit model that I discuss in the next section descend from Yule's and Pearson's different measures of association. One interpretation of the logit model is epistemically conservative, whereas the other posits the existence of an underlying latent variable that determines the binary outcome variable. This interpretation underlies the use of the logit model to measure unobserved attributes in psychology and economics.

3.4. Two approaches to the logit model

The outputs of a logistic regression model are most straightforwardly interpreted as the log-odds of a particular outcome: $\log\{[P(Y=1)]/[1-P(Y=1)]\}$. This outcome can be easily converted to the odds of a particular outcome (by exponentiating) or a probability using the sigmoid function:

https://doi.org/10.1017/psa.2024.25 Published online by Cambridge University Press

$$s(x) = \frac{1}{1 + \exp(-x)}. \tag{8}$$

Although the logit model estimates, just like Yule's Q association measure, are computed from observed frequencies, there is a further question of how the output should be interpreted. Literally, the output of a logistic regression is a log-odds. But does the log-odds reflect an underlying continuous variable that decides category membership in the same way that Pearson claims that mortality reflects the severity of a disease? If so, then logistic regression can be seen as a model involving hidden variables—a latent variable model. Alternatively, does the log-odds summarize the expected long-run frequencies? Or does it represent a logical connection between the information obtained from predictor variables and the outcome variable? If so, the model is fundamentally probabilistic. The various approaches to the logit model discussed in this article differ with regard to how they answer the question of how to interpret the output of the logit model.

3.4.1. The transformational approach

The transformational approach to the logit model interprets the outcome variable as a log-odds of success ($Y = 1$). This is an estimate of an observable quantity, because the actual odds of success can be computed from the observed proportions (Powers and Xie 2008, 32).

The transformational approach to the logit model is a descendant of Yule's approach to measuring association between binary variables. Yule's measure transformed observed binary data into proportions and odds. All the quantities involved in Yule's measure correspond to empirical quantities: quantities that are recorded in the data set.

Consider a logit model that predicts the odds of a convicted criminal reoffending given the number of past convictions. According to the transformational approach, the outcome variable of this model is a function of a predictive probability. It can be directly compared with the empirical probability (or some function of it), the actual observed odds of a convicted criminal reoffending given the number of past convictions.

Perhaps some underlying unobserved attribute is causally responsible for the observed binary outcome and observed probabilities. Although it does not represent such unobserved attributes, the transformational approach is not inconsistent with the existence of one or multiple latent variables that explain the observed probabilities. Whether or not these unobserved attributes explaining the binary outcome of the logit model exist, the logit model is restricted to modeling quantities that correspond to observed data.

3.4.2 The latent variable approach

The latent variable approach to the logit model defines the outcome variable as a representation of an unobserved attribute that is not reflected in the data. Consider again the logit model that predicts the odds of a convicted criminal reoffending given the number of past convictions. Instead of interpreting the outcome variable as a function of the predictive probability, the latent variable approach legitimizes the interpretation of the outcome variable of the model as a distinct attribute that

https://doi.org/10.1017/psa.2024.25 Published online by Cambridge University Press

explains the observed frequencies, probabilities, and odds. For example, the latent variable could be recidivism, the tendency of a criminal to reoffend.

It is generally accepted that this approach to the logit model is a descendant of Pearson's measure of association based on an underlying continuous variable (Agresti 2002; Breen, Karlson, and Holm 2018; Cox and Snell 1989; Powers and Xie 2008). A continuous underlying variable Z is posited. At a particular threshold of Z, the observed binary outcome changes from 1 to 0. If, for example, the observed variable is state of matter and the underlying variable is temperature, at a threshold 32°F, the observed variable changes because of the value of the underlying variable.

Mathematically, the latent variable approach to the logit model is expressed as follows. Suppose that Z is a latent variable,

$$Z_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_j x_{ij} + \epsilon_i, \tag{9}$$

and

$$\text{if } Z_i \geq 0, Y_i = 1 \tag{10}$$

$$\text{if } Z_i < 0, Y_i = 0. \tag{11}$$

The error term ϵ_i is assumed to follow a standard logistic distribution with a fixed variance. Because the model is not considered a probability model in this case, it is necessary to express the indeterminacy of the model in an error term. That the variance of the error is fixed will play an important role in my subsequent argument that the outcome of a logit model cannot be interpreted as a measurement of a latent variable. From this distributional assumption, it follows that

$$\text{logit}(P(Y = 1)) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_j x_{ji}. \tag{12}$$

Note that unlike the transformational approach, the latent variable approach includes an error term. The outcome variable in the latent variable approach is not considered to be fundamentally probabilistic; rather, it is taken to represent the value of a random variable that is unobserved. Thus the uncertainty is not captured in the outcome variable of the model. To capture the uncertainty inherent in a regression model, then, an error term is tacked on. The assumption that the error term follows a standard logistic distribution defines the relationship between the logit and the unobserved random variable.

The central question of this article is, do statistical models yield measurements? In the social sciences, statistical models are sometimes treated as measurement instruments, meaning that the output of these models are measurements. The latent variable approach to logistic regression is central to the use of statistical models as measurement instruments in the social sciences. According to the latent variable approach, underlying attributes can be represented by statistical models of observed binary data. In fields that aim to study unobservables like psychological traits or economic value, this interpretation of the logit model suggests a way to access the attributes being studied (Everitt 1984; Reise, Ainseoth, and Haviland 2005).

Discrete choice theory in economics, for example, makes use of the latent variable interpretation of the logit model to measure attributes of individuals that compel them to make particular choices. Daniel McFadden (1986, 275), the founder of discrete

https://doi.org/10.1017/psa.2024.25 Published online by Cambridge University Press

choice theory, describes it as a way to "model explicitly the cognitive mechanisms in the black box that govern behavior."

The latent variable interpretation of the logit model also plays a central role in psychometrics, the study of the measurement of psychological traits (Borsboom, Mellenbergh, and Van Heerden 2003; Borgstede and Eggert 2023; Hood 2013). The Rasch model is a psychometric model that is a special case of the two-parameter logistic model (Anderson 2009; Rasch 1960). Statistician Carolyn Anderson (2009, 325) describes the psychometric framework as assuming "a latent continuum" that "may be utility, preference, skill, achievement, ability, excellence, or some other construct" underlying binary test data.

3.4.3. Reflecting on the two approaches

To summarize, the latent variable approach to the logit model can be traced back to a view of binary data held by Karl Pearson. Pearson asserted that binary data arise from an underlying continuous variable. Statistical modeling can be used, Pearson claimed, to reconstruct that underlying variable from observed data. This interpretation of the logit model is widely in use today, in particular in the social sciences. In fields like psychology and economics, the logit model is used as a measurement instrument, providing indirect access to unobservable attributes via binary test data.

The transformational approach, on the other hand, treats the logit model, not as representative of a hidden quantity, but rather as fundamentally probabilistic. On this view, the logit model could still be considered a measurement instrument, but not of an unobservable attribute; rather, it measures a probability.[4] Such a view may be possible, but it would not support use of the model to measure hidden variables in the social sciences.

The argument between Yule and Pearson can help us understand why the logit model output is sometimes seen as a measurement of a hidden variable. To decide whether it's appropriate to regard it as a measurement, however, we need a theory of what a measurement is. So, in the next section, I explore philosophical theories of measurement.

4. The measurement question

In the previous section, I showed that the latent variable approach to the logit model championed by Karl Pearson provides a historical explanation for why the logit model is treated as a measurement instrument in the social sciences. In this section, I turn to the question of whether the logit model is a genuine measurement instrument.

First, I argue that measurement instruments necessarily produce comparable outcomes.[5] If outcomes produced by a measurement procedure cannot be meaningfully compared, then it can be concluded that genuine measurement has not taken place. I show that this criterion for measurement follows from the view that measurement maps empirical relations to numerical relations. Then, I show that the comparability criterion possesses the following attractive quality: it is consistent with a variety of philosophical positions on measurement. I show that both foundationalists and antifoundationalists about measurement can adopt my criterion without

[4] Thanks to an anonymous reviewer for pointing this out.

[5] In this section, I use *measurement instrument* interchangeably with *measurement procedure*.

https://doi.org/10.1017/psa.2024.25 Published online by Cambridge University Press

challenge to their respective position. Finally, I show that the logit model violates comparability.

4.1. The comparability constraint

Some procedures can be ruled out as measurement procedures from the get-go. We can, for example, rule out the following physical procedure for measuring one's height: Take the nearest physical object (no need to report what it is) and count how many times this object can lie end to end from head to toe. Report the number of times. The following inferential procedure for measurement can be ruled out too: Take an instrument indication and multiply it by a random number between 1 and 10. Report that magnitude as your result.

Why do these procedures fail to yield measurements? In each case, the procedure violates a criterion I call *comparability*. Consider the following: if you report your height to be ten using the "nearest physical object" method, and I report mine to be nine, I am still in the dark about the empirical relationship between my height and yours. Comparability is the preservation of empirical relations between quantities in the numerical outcome of measurement. In the case of height, the empirical relations "shorter than" and "taller than" should be preserved by the measurement procedure for comparability.

In this section, I motivate the adoption of the *comparability criterion for measurement*: if a procedure violates comparability, then it is not a measurement. In other words, if no empirical relations between the quantities that are the subject of measurement are preserved in the numerical outcomes given by a procedure, then that procedure is not a measurement. Comparability, then, is a necessary condition of measurement.

The representational theory of measurement (Suppes and Zinnes 1963; Krantz, Suppes, and Luce 1989; Luce et al. 1990; Luce and Suppes 2002) proves to be a useful formalization for my purposes—comparability can be seen to follow from the use of a representational theorem to map empirical quantities to numerical scales that preserve order.[6] According to the representational theory, there are two central problems for measurement: (1) what is the representation theorem that maps empirical relations to numerical relations (finding a representation theorem)? and (2) given a representational theorem, what transformations preserve the relations between the representations (finding a uniqueness theorem)?

The representation theorem maps empirical relations to numerical values. For example, a representation theorem for assigning numbers to lengths of rods will map each rod to the real numbers, the operation of concatenation (the action of linking things together) to addition, and the comparison of physical rods (shorter than, longer than) to the comparison of real numbers (less than, greater than).

The uniqueness theorem defines the transformations that preserve the relations given by the representation theorem. Some scales permit more meaningful transformations than others. For example, a ratio scale can be multiplied and

[6] Here I will use some uncontroversial mathematical ideas from the representational theory of measurement to spell out a notion in which I am interested: comparability of outcomes. Many are critical of the representational theory for its inadequate treatment of the epistemic and ontological problem of measurement (Heilmann 2015; Michell 2021; Tal 2021). I set aside these issues for the purposes of this discussion.

https://doi.org/10.1017/psa.2024.25 Published online by Cambridge University Press

divided, whereas an interval scale cannot be. A ratio scale is a scale with a true zero, such as length. The zero point on the scale indicates an actual absence of length. I can compare length multiplicatively—it is meaningful to claim that the distance from my house to the university is twice the distance from my house to the grocery store. This holds true regardless of whether I measure the distance in meters or yards. Interval scales have an arbitrary zero, such as temperature. Temperatures cannot be compared multiplicatively—it is not meaningful to claim that it is twice as hot today as yesterday, because this relation is not preserved across scales with different zero points (e.g., Fahrenheit vs. Celsius). Ordinal scales indicate order but violate additivity. Ratio and interval scales distribute numerals evenly, such that there is the same distance between 1 and 2 and 3 and 4. Ordinal scales make no such guarantee as to the distance between each numeral, and so additivity is violated.

The representation theorem preserves empirical relations present in the target system: suppose the representation theorem maps the empirical relations of shorter than ($\prec$) and longer than ($\succ$) to the numerical relations $<$ and $>$; then, the representation theorem must be such that if $a \prec b$, then R is a representation theorem only if $R(\prec, a, b) \rightarrow A < B$. The empirical relations preserved over and above mere ordering (like additivity) determine to which type of scale the empirical relations are mapped.

Comparability is implied when empirical relations are mapped to any scale that preserves order: ordinal, interval, or ratio.[7] Note that comparability is not a formal property of a scale—instead, it is a property of measurement outcomes that is guaranteed by the successful mapping of empirical relations to an ordinal, interval, or ratio scale.

Recall the nonmeasurement procedures described earlier—the "nearest physical object" method of measuring height, for example. What goes wrong in this case is that the "representation theorem" employed fails to preserve any empirical relations in the mapping to numerical relations. Because two individuals who use this method will have different units for measure (depending on the physical object nearest to them), it is not guaranteed that given $a \prec b$, $R(\prec, a, b) \rightarrow A < B$.

What it is to measure is to map empirical relations onto an ordinal, interval, or ratio scale. A measurement procedure performs that mapping while preserving the empirical relations present in the quantity of interest. If what it is to measure is to map empirical relations onto an ordinal, interval, or ratio scale while preserving the empirical relations that are present in the quantity of interest, then comparability is entailed by measurement.[8] A lack of comparability in measurement outcomes, then, indicates that measurement has not taken place.

[7] Nominal scales do not preserve order. Although Stevens (1946) includes nominal scales in his theory of measurement, nominal scales are not usually considered a type of measurement (Luce and Suppes 2002, 14). My comparability criterion rules out nominal scales for measurement, but this is uncontroversial.

[8] Recall that comparability is the preservation of empirical relations between quantities in the numerical outcome of measurement.

https://doi.org/10.1017/psa.2024.25 Published online by Cambridge University Press

4.2. Comparability and the foundationalism debates

An attractive feature of the comparability criterion that I propose is that it is relativizable to a theory and thus can be adopted by both foundationalists and antifoundationalists about measurement without challenge to their respective view.

Foundationalism about measurement is the view that measurements reflect pretheoretical observations that form the foundation for scientific theorizing (Thalos 2023, 1). Antifoundationalism about measurement is the view that measurements are dependent on theoretical assumptions (Chang 2004; Tal 2012; van Fraassen 2012). Contemporary discussion of psychological measurement is often focused on this debate (see Michell 1993, 1997, 1999, 2005; Hood 2013). In this article, though, I offer a critique of a method employed in psychological measurement that sidesteps this contentious issue.

Foundationalism is consistent with comparability as a criterion for measurement because the foundationalist is committed to the view that the use of ordered scales in measurement reflects real quantitative properties possessed by the measurement subject. Given this, it is necessary for the measurement procedure, according to the foundationalist, to accurately map those empirical relations to numerical relations. I have argued that comparability follows from this view.

Antifoundationalist accounts initially appear to be in greater tension with the adoption of the comparability criterion than foundationalist accounts are. According to antifoundationalist views like the model-based theory of measurement advanced by Eran Tal, measurements are made relative to an idealized model of the measurement process. This model formalizes background theoretical assumptions and statistical assumptions about error. For example, in measuring temperature, the measurement depends upon broad background theoretical assumptions (a physical theory of temperature) and also statistical assumptions about the error expected given slight variations in the conditions under which the measurement takes place (Tal 2019, 864). Measurement, then, is theory laden, meaning it cannot be done independently of assumptions about the quantity being measured and the process by which the quantity is measured.

There are constraints on which models yield measurement outcomes: measurements must be based on models that are coherent and that produce outcomes that are "consistent with other measurements of the same or related general quantity types performed under different conditions" (Tal 2019, 871). A model is coherent when it coheres with the physical theories that are presupposed prior to measurement. A model is consistent when outcomes of different measurement procedures agree within an acceptable margin of error. Consistency of a model establishes the *objectivity* of a measurement, whereby a measurement is objective when it is "about measured objects rather than about features of the measurement operation" (Tal 2019, 866). Tal goes on to state that "without this minimal sort of objectivity, measurement itself is impossible" (866).

Objectivity is model relative, however, meaning that objective measurements can be made only after some initial background assumptions have been accepted. For example, the kinetic theory of matter serves as a theoretical background that justifies the use of thermometers to measure temperature. Our measurements of temperature

https://doi.org/10.1017/psa.2024.25 Published online by Cambridge University Press

tell us about the measured object in the world, but always with respect to our background theories.

Like objectivity, comparability can be model relativized in the following way: model-relative comparability is the preservation of empirical relations between quantities in the numerical outcome of measurement given the relevant background theories. So, even when measurements cannot be taken independently of theoretical assumptions, measurements can be deemed comparable given the theoretical assumptions that ground the possibility of measurement.

4.3. Logit model measurement

In this section, I describe how the logit model is used for measurement in psychometrics. Then, I argue that the logit model is not a genuine measurement instrument of a latent trait because it violates the comparability criterion for measurement argued for earlier.

4.3.1. The Rasch model

The latent variable approach to the logit model is used in psychometrics to measure unobservable psychological traits. The Rasch model (developed by Georg Rasch), for example, is a special case of the logit model that is used to measure ability or some other latent trait given dichotomous test data.

The Rasch model takes the probability of correctly answering a question as a function of difficulty and ability (or some other latent trait); that is,

$$P(\text{Correct}) = \frac{1}{1 + e^{-(\text{ability} - \text{difficulty})}}. \tag{13}$$

This expression is equivalent to a logistic regression with two predictor variables, ability and difficulty. In the context of a Rasch model, the ability coefficient is interpreted as a measure of a latent trait.

In what follows, I argue that the logit model lacks comparability and thus cannot be used for measurement. This would challenge the use of logit models like the Rasch model for psychological measurement.

4.3.2. The logit model and comparability

Earlier, I explained that the logit model is a regression model for binary outcome variables. The logit model can be interpreted as fundamentally probabilistic or as a representation of a latent variable. When it is interpreted as a latent variable model, the model includes an error term with a standard logistic distribution. In what follows, I (1) review the role of the error term in the logistic regression, (2) explain why the logit model has a fixed error variance, and (3) show that the fixed error variance results in the following property: the model coefficients are altered whenever a new predictor is added to the model or the model is used in a new context. Then, I argue that this property violates the comparability criterion for measurement.

4.3.2.1. The error term

A regression partially explains the variation observed in the outcome variable with predictor variables. For example, yearly salary varies in a population, meaning that not everyone has the same salary. A regression model that predicts yearly salary

https://doi.org/10.1017/psa.2024.25 Published online by Cambridge University Press

given years of education explains some of the population variation in salary by years of education; that is, if the population were to be stratified by years of education, there would be less variation in each stratum than in the population as a whole. Because regression models are indeterminate, not all of the variation in the outcome variable is explained by the predictor variables.

The variation in the outcome variable that is left unexplained by a model is called the *unexplained variance*. Adding more predictor variables to the model will increase the amount of explained variance and decrease the unexplained variance. The unexplained variance is captured in the model by the error term ϵ. The error term covers the difference between the outcome predicted by the model and the true outcome for each observation. The error term has its own distribution that characterizes how the model errors are distributed. Ideally, it should have a distribution with a mean of 0 (indicating that the model predictions are not systematically high or low) and a low variance (indicating that the model makes accurate predictions).

4.3.2.2. Fixed error variance

In both linear and logistic regression, the distribution of the error term ϵ follows from the assumed distribution of the outcome variables. In the case of a linear regression, ϵ is normally distributed. In the case of a logistic regression, ϵ has a logistic distribution. The error term in linear regression is assumed to have a mean of 0 and a variance that is estimated from the data. The variance of the error term in a linear regression is estimated by calculating the difference between the predicted and actual values for each observation and finding the variance of these differences. The variance of the error term in logistic regression, though, cannot be estimated from the data. The latent variable approach interprets the model as estimating a value for a continuous underlying variable, but the actual value of this variable is unobserved. The difference between the prediction and actual values cannot be calculated.

To estimate the parameter values of the model, the variance of the errors must be assumed to have a fixed value because binary data have no "inherent scale" (Powers and Xie 2008, 58).[9] Fixing the value of the variance of the errors serves to set a unit for the model. The logit model error term is assumed to have a standard logistic distribution with a fixed error term of $\pi^2/3$.

As a consequence of fixing the variance of the errors, the logit model parameter estimates are scaled by an unknown quantity that depends on the actual variance of the errors. The true variance of the error term is unknown and unknowable, because the only way of observing this variance would be to observe the true values of the latent variable that is represented by the logit model (Breen, Karlson, and Holm 2018, 41).

4.3.2.3. Scaled coefficients

One important consequence of the fixed error variance of the logit model is that coefficient estimates are altered whenever a new predictor variable is added to the

[9] In this section, I will be using the word *scale* to mean the unit of measure or (as a verb) to change the unit of measure. This is consistent with the word's use in statistical contexts but different from its use in measurement contexts.

https://doi.org/10.1017/psa.2024.25 Published online by Cambridge University Press

model, regardless of whether this new predictor variable is correlated with the predictor variables already in the model (Mood 2010). This is highly unintuitive.

To appreciate how surprising this property is, consider a *linear* regression model that predicts student performance on an exam by hours of studying. A coefficient is estimated and interpreted to mean that, on average, for every additional hour of studying, there is a 5-point increase in exam score. Now, imagine that we add a predictor to our model: hours of sleep the night before the exam. If this new predictor is uncorrelated with hours of studying, we don't expect the relationship between hours of studying and exam score to change, because the new predictor doesn't "steal" any of the explained variance from our old predictor. If, on the other hand, the new predictor is correlated with hours of studying, then we expect the strength of the relationship between hours of studying and exam score to decrease. Equivalently, the coefficient estimate will decrease. Some of the variance that was explained by the old predictor in the old model is taken by the new predictor in the new model. For linear regression, any change in the coefficient estimates upon the addition of a new variable to the model can be interpreted as reflecting correlation between the predictor variables.

The logit model has a fixed error variance. Although adding new uncorrelated predictors to the model will increase the amount of explained variance, it cannot be compensated for by a smaller error variance, which must remain $\pi^2/3$. Instead, all model terms are scaled differently, meaning that the coefficients for predictor variables already in the model are altered.

To understand the effect, consider the following abstracted example. Imagine that we are interested in modeling a variable Y using a predictor variable X_1 and that the total amount of variance in our data of Y is 10. First, suppose our model is a linear regression. The regression splits up the variance in Y such that 5 is explained by the predictor variable X_1 and 5 remains unexplained, meaning the variance of the errors is 5.

Now, suppose another predictor, X_2, is added to the model. If X_2 is uncorrelated with X_1, then the variance of Y could be split in the following way: X_1 still explains 5, X_2 explains 3, and 2 remains unexplained, meaning that the variance of the errors in this model is 2. If X_2 is correlated with X_1, then the variance of Y could be split in the following way: X_1 now explains 4, X_2 explains 4, and 2 remains unexplained, meaning that the variance of the errors in this model is 2. Observe that when the predictor variables were correlated in the abstracted case, the second predictor variable "stole" explained variance from the first predictor variable.

Now, let's imagine that our model is a logistic regression with a fixed error variance. Whatever the actual amount of unexplained variance is, it is called $\pi^2/3$ (or approximately 3.29) in our model. Now, imagine that an uncorrelated predictor variable is added to the model. Although this decreases the amount of unexplained variance, it is still called $\pi^2/3$ in our model. That means that the unit has changed, because a different amount of variance is labeled with the same numerical value. A new unit is fixed for this model whenever the amount of unexplained variance changes. The coefficients for the model are then estimated using this different unit, meaning that although the actual amount of variance explained by a predictor is unchanged, the coefficient estimate will change. This is why, even when new

https://doi.org/10.1017/psa.2024.25 Published online by Cambridge University Press

predictors are added to the model that are uncorrelated with predictors already in the model, the coefficient estimates will change.

Coefficients cannot be treated as straightforward effects. They reflect not only the effect of the predictor variable on the outcome variable but also some unknown scale quantity. Whenever the amount of unexplained variance in the model changes, the scale quantity changes too. Coefficients across models with different sets of predictor variables cannot be compared, because the amount of unexplained variance is different across models with different sets of predictors (Karlson, Holm, and Breen 2012). So, if two logit models are constructed to target the same latent trait (such as well-being) with different predictor variables, the models will produce results given in different units.

Similarly, outcomes of models with the same predictor variables on different samples cannot be compared, because the variance of the outcome variable is guaranteed to be different from sample to sample (Breen, Karlson, and Holm 2018). So, if one logit model is used to target a latent trait on two different samples (perhaps two individuals), the model will produce results that are given in different units. The relation between the units is unknown, because the true amount of unexplained variance is unknown. This is analogous to the "nearest physical object" method of measuring height—the procedure fails to preserve the empirical relations in the mapping to numerical relations. Logit models are single use: they provide information about the relative importance of predictor variables only within a single model applied to a single sample.

Previously, I suggested that comparability is a necessary condition of measurement. If the empirical relations between quantities are not preserved in the numerical outcome of a procedure (relative to relevant background theories), then it is not a measurement. I have shown that outcomes of the logit model are given in different units with an unknown relationship, and as such, the empirical relationship between different outcomes is not preserved by the procedure. The logit model produces incomparable outcomes and thus is not a measurement instrument.[10]

Statisticians have periodically highlighted the properties of logistic regression that I have described, although the consequences of these properties for the use of logistic regression as a measurement instrument in the social sciences have not been considered (Gail, Wieand, and Piantadosi 1984). Mood wrote in 2010 that "these insights have not penetrated our research practice and the inter-relations between the problems have not been fully appreciated. Moreover, these problems are ignored or even misreported in commonly used methodology books" (68). Nearly ten years later, in 2018, Breen, Karlson, and Holm made the same observation: "empirical

[10] It could be replied that although the coefficients are scaled by an unknown factor, the values of the coefficients within a single model can still be used to assess the relative importance of coefficients in the model in predicting the outcome.

It is correct that the logit model coefficient estimates could impose an ordering of importance over the predictor variable included in the model. This would constitute an ordinal scale, but not an ordinal scale over the attribute of interest. In using logistic regression as a measure of a latent variable, the practitioner is interested in creating a scale of the latent variable, not over the different predictor variables included in the model.

The following reply could also be made: a priori consideration of the logit model form is unnecessary given methods to establish validity and reliability of psychometric scales. I leave this for future work.

https://doi.org/10.1017/psa.2024.25 Published online by Cambridge University Press

applications of these nonlinear probability models seldom take account of a body of work that, over the past three decades, has pointed to their problematic aspects and, particularly, to difficulties in interpreting their parameters" (40). The relative obscurity of this property of the logit model may explain why it hasn't been addressed in the body of work in psychometrics that treats logistic regression as a measurement instrument.

5. Conclusion

Do statistical models yield measurements? I've argued in this article that the logit model, a particular statistical model, does not yield measurements of latent traits despite how it is used in the social sciences.[11] Comparability, I argued, is a constraint on what counts as a measurement instrument that is not satisfied by the logit model. This result depends on the particular properties of the logit model and thus doesn't generalize to all statistical models.

In this article, I hope to have demonstrated the following metaphilosophical principle: statistical models should be evaluated individually to determine the meaning of their outputs. Social scientists have developed applied methods to cope with the circumstances in their fields, and in doing so, they have neglected the task of establishing that the instruments in use preserve the empirical relations of interest in the mapping to numerical relations.

Acknowledgements. Thanks to Sinan Dogramaci, Sahotra Sarkar, Christian Hennig, Jim Hankinson, and the organizers and participants of the 2023 Measuring the Human Workshop for comments and discussion. Thanks especially to the reviewers for their excellent suggestions that improved the article.

References

Agresti, A. 2002. *Categorical Data Analysis* (2nd ed.). Hoboken, NJ: Wiley-Interscience. https://doi.org/10.1002/0471249688.

Anderson, C. J. 2009. "Categorical Data Analysis with a Psychometric Twist." In *The SAGE Handbook of Quantitative Methods in Psychology*, edited by R. Millsap and A. Maydeu-Olivares, 311–36. Thousand Oaks, CA: SAGE. https://doi.org/10.4135/9780857020994.n14.

Borgstede, M., and F. Eggert. 2023. "Squaring the Circle: From Latent Variables to Theory-Based Measurement." *Theory and Psychology* 33 (1):118–37. https://doi.org/10.1177/09593543221127985.

Borsboom, D., G. J. Mellenbergh, and J. Van Heerden. 2003. "The Theoretical Status of Latent Variables." *Psychological Review* 110 (2):203–19. https://doi.org/10.1037/0033-295X.110.2.203.

Breen, R., K. B. Karlson, and A. Holm. 2018. "Interpreting and Understanding Logits, Probits, and Other Nonlinear Probability Models." *Annual Review of Sociology* 44 (1):39–54. https://doi.org/10.1146/annurev-soc-073117-041429.

Chang, H. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press. https://doi.org/10.1093/0195171276.001.0001.

Cox, D. R., and E. J. Snell. 1989. *Analysis of Binary Data* (2nd ed.). New York: Chapman and Hall.

Everitt, B. S. 1984. *An Introduction to Latent Variable Models*. Dordrecht, Netherlands: Springer. https://doi.org/10.1007/978-94-009-5564-6.

[11] An anonymous reviewer pointed out to me that it may be possible to conceive of the logit model as yielding measurements of a probability, even if it is not yielding measurements of a latent variable. I will flag here that it may be possible to argue that the logit model measures *epistemic* probability, where epistemic probability is fundamentally conditional on the available evidence and thus comparability across models with different evidence is not required. Lacking the space to develop this idea here, I will leave such a possibility for future work.

https://doi.org/10.1017/psa.2024.25 Published online by Cambridge University Press

Gail, M. H., S. Wieand, and S. Piantadosi. 1984. "Biased Estimates of Treatment Effect in Randomized Experiments with Nonlinear Regressions and Omitted Covariates." *Biometrika* 71 (3):431–44. https://doi.org/10.2307/2336553.

Heilmann, C. 2015. "A New Interpretation of the Representational Theory of Measurement." *Philosophy of Science* 82 (5):787–97. https://doi.org/10.1086/683280.

Hood, S. B. 2013. "Psychological Measurement and Methodological Realism." *Erkenntnis* 78 (4):739–61. https://doi.org/10.1007/s10670-013-9502-z.

Karlson, K. B., A. Holm, and R. Breen. 2012. "Comparing Regression Coefficients Between Same-Sample Nested Models Using Logit and Probit: A New Method." *Sociological Methodology* 42 (1):286–313. https://doi.org/10.1177/0081175012444861.

Krantz, David, Patrick Suppes, and Robert D. Luce. 1989. *Foundations of Measurement: Geometrical, Threshold, and Probabilistic Representations*. New York: Academic Press.

Luce, R. D., D. H. Krantz, Patrick Suppes, and A. Tversky. 1990. *Foundations of Measurement: Representation, Axiomatization, and Invariance*. New York: Academic Press. https://doi.org/10.1016/B978-0-12-425403-9.50010-2.

Luce, Robert D., and Patrick Suppes. 2002. "Representational Measurement Theory." In *Stevens' Handbook of Experimental Psychology: Methodology in Experimental Psychology*, vol. 4, 3rd ed., edited by D. Luce and R. Bush, 1–41. Hoboken, NJ: John Wiley. https://doi.org/10.1002/0471214426.pas0401.

McFadden, D. 1986. "The Choice Theory Approach to Market Research." *Marketing Science* 5 (4):275–97. https://doi.org/10.1287/mksc.5.4.275.

Michell, J. 1993. "The Origins of the Representational Theory of Measurement: Helmholtz, Hölder, and Russell." *Studies in History and Philosophy of Science, Part A* 24 (2):185–206. https://doi.org/10.1016/0039-3681(93)90045-L.

Michell, J. 1997. "Quantitative Science and the Definition of Measurement in Psychology." *British Journal of Psychology* 88 (3):355–83. https://doi.org/10.1111/j.2044-8295.1997.tb02641.x.

Michell, J. 1999. *Measurement in Psychology: A Critical History of a Methodological Concept*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511490040.

Michell, J. 2005. "The Logic of Measurement: A Realist Overview." *Measurement* 38 (4):285–94. https://doi.org/10.1016/j.measurement.2005.09.004.

Michell, J. 2021. "Representational Measurement Theory: Is Its Number Up?" *Theory and Psychology* 31 (1):3–23. https://doi.org/10.1177/0959354320930817.

Mood, C. 2010. "Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do about It." *European Sociological Review* 26 (1):67–82. https://doi.org/10.1093/esr/jcp006.

Pearson, K. 1904. *On the Theory of Contingency and Its Relation to Association and Normal Correlation*. Biometric Series. London: Dulau.

Pearson, K., and D. Heron. 1913. "On Theories of Association." *Biometrika* 9 (1–2):159–315.

Powers, D., and Y. Xie. 2008. *Statistical Methods for Categorical Data Analysis*. New York: Emerald. https://doi.org/10.1111/j.1751-5823.2010.00118_3.x.

Rasch, G. 1960. *Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests*. London: Nielsen and Lydiche.

Reise, S., A. Ainseoth, and M. Haviland. 2005. "Item Response Theory: Fundamentals, Applications, and Promise in Psychological Research." *Current Directions in Psychological Science* 14 (2):95–101. https://doi.org/10.1111/j.0963-7214.2005.00342.x.

Stevens, S. S. 1946. "On the Theory of Scales of Measurement." *Science* 103 (2684):677–80. https://doi.org/10.1126/science.103.2684.677.

Suppes, P., and J. Zinnes. 1963. "Basic Measurement Theory." In *Handbook of Mathematical Psychology*, vol. 1, edited by D. Luce and R. Bush, 1–76. Hoboken, NJ: John Wiley. https://doi.org/10.2307/2270274.

Tal, E. 2012. "The Epistemology of Measurement: A Model-Based Account." PhD diss., University of Toronto.

Tal, E. 2019. "Individuating Qualities." *Philosophical Studies* 176 (4):853–78. https://doi.org/10.1111/phc3.12089.

Tal, E. 2021. "Two Myths of Representational Measurement." *Perspectives on Science* 29 (6):701–41. https://doi.org/10.1162/posc_a_00391.

Thalos, M. 2023. "The Logic of Measurement: A Defense of Foundationalist Empiricism." *Episteme*. https://doi.org/10.1017/epi.2023.32.

https://doi.org/10.1017/psa.2024.25 Published online by Cambridge University Press

van Fraassen, B. 2012. "Modeling and Measurement: The Criterion of Empirical Grounding." *Philosophy of Science* 79 (5):773–84. https://doi.org/10.1086/667847.

Yule, G. U. 1900. "On the Association of Attributes in Statistics, with Examples from the Material of the Childhood Society." *Philosophical Transactions of the Royal Society of London, Series A* 194:257–319. https://doi.org/10.1098/rsta.1900.0019.

Yule, G. U. 1912. "On the Methods of Measuring Association between Two Attributes." *Journal of the Royal Statistical Society* 75 (6):579–652. https://doi.org/10.1111/j.2397-2335.1912.tb00463.x.

Cite this article: Fillmore-Patrick, Stella. 2025. "The Logit Model Measurement Problem." *Philosophy of Science* 92 (2):285–303. https://doi.org/10.1017/psa.2024.25