Construct Validity and Reliability of the IRIS Dataset

This paper will argue that the construct validity and reliability of the IRIS dataset are compromised by, among other things, the subjective data collection procedure employed, a poorly-defined systematized concept, subjective measurement strategy, restricted generalizability across constructs, inadequate preoperational explication of constructs, failure to properly operationalize particular concepts, bias in experimenter expectancies, and inherent weaknesses stemming from the original purpose for which the data were collected.

The IRIS dataset was originally developed in 1993 by Philip Keefer and Stephen Knack in their work for the Institutional Reform and the Informal Sector (IRIS) research and advisory center at the University of Maryland. The dataset is based on data obtained from the International Country Risk Guide (ICRG), and contains scores for six "political risk variables," comprising corruption, the rule of law, bureaucratic quality, ethnic tensions, repudiation of government contracts, and expropriation risk. The latest version of the IRIS dataset (IRIS-3) covers the years from 1982 to 1997. IRIS is based on proprietary institutional indicators compiled by ICRG.

It is important to stress that these data are not being used for the purpose for which they were originally compiled. ICRG assesses potential risks to international business operations, and the assessments are made by its staff using a mix of subjective and objective criteria (ICRG Methodology, PRS 2003). The focus on international investment raises questions about the predictive power of the data in other contexts. In addition, it is unclear whether the data are interval or ordinal measures – we would like to know whether a 4 actually twice as good as a 2, for instance. It is difficult to know whether the data are measuring all of the key attributes of good governance without some "extra" factor.

Construct validity refers to the degree to which inferences can legitimately be made from operationalizations to the theoretical constructs on which those operationalizations are based (Trochim and Donnelly 2007). To properly assess the construct validity of the IRIS dataset, we will need to examine what the data are purporting to measure. The construct in the IRIS dataset is "quality of governance," which is a composite measure (ICRG82) of six different subsidiary constructs, each operationalized by a measure in the dataset. Knack's (1995) loosely-defined systematized concept is "property rights," and he claims that the IRIS dataset tracks the particular institutions of government emphasized by Douglass North (1990), Barry Weingast

(1993), and Mancur Olson (1982). Knack claims that the data are predictive of private rates of investment, which relates to the criticism outlined above: that these data were collected to measure investment risk arising from political factors (Knack 1995). It seems that the data do indeed measure "property rights," but with a strong emphasis on the perspective of international investment. This shortcoming renders the data less predictive in domestic political contexts.

We will first examine the six measures set out in the IRIS dataset. To assess the validity of Knack's operationalization, we will ask whether each measure produces scores that adequately capture the systematized concept. First, we will examine translation validity. We would like to know whether the six IRIS measures enumerate the key components of the construct and nothing else. We can separate this question into face validity and content validity, and we will examine each of the measure in turn.

Quality of the bureaucracy – This operationalization has high face validity, as it appears that a high-quality bureaucracy would indeed be independent from political pressure and consistent in its objectives. However, when examining its content validity the operationalization exhibits several shortcomings. A high-quality bureaucracy would (additionally) exhibit such features as responsiveness, impartiality and incorruptibility, none of which are adequately captured by this operationalization.

Corruption in government – The face validity of this operationalization is more questionable, as it fails to capture grand corruption (by a state's rulers) or the bribes necessary to achieve a citizen's goals in a corrupt context (such as bribes for utility connection, job placement, or children's education). The content validity of the measure is even more troubling, as the relevant content domain must, at a minimum, include politicians as well as officials and reflect the experience of citizens, not merely corporations doing business in the state concerned.

Rule of law – The face validity of this operationalization is limited by the different criteria for high and low scores. The high scores are based on an objective measures – institutions and laws – while the low scores are subjective, based on a tradition of using force or illegality. Similarly, its content validity is undermined by a change in the framing of the construct. A state's "tradition of law and order" is not the same as the "rule of law" – to state the obvious, a despotic regime may exhibit a high degree of law and order without significant rule of law. Redefining the construct as "rule of law" adds an important normative component (Davis, Fisher et. al. 2012). In addition, this operationalization seems to violate the "and nothing else"

injunction – combining orderly transfer of power with citizens' propensity to obey the established laws appears to conflate two distinct elements.

Ethnic tensions – The face validity of this operationalization is unproblematic – it sets out virtually a dictionary definition of ethnic tensions. An immediate question when assessing content validity, however, is the reliability of the mechanism for categorizing particular tensions as "ethnic" or otherwise. As this aspect of the methodology is proprietary and performed by ICRG staff, further analysis is difficult, but the inherent subjectivity of the measure is worrying.

Expropriation risk – The face validity of this operationalization is questionable – what sort of outright confiscation are we talking about? Does it apply equally to domestic and foreign property owners? The content validity fares no better. An objective list of the criteria for expropriation risk would need to specify the type of property at risk of expropriation (e.g. capital? Land? Corveé labor?) as well as the type of property owner subject to it.

Repudiation of contracts by government – Leaving aside whether this operationalization is in fact a form of expropriation, the definition adequately guarantees face validity – the risk of a modification to a contract is more or less the risk of repudiation of contracts, though the definition leaves out "involuntary," which seems vital for a valid operationalization. Construct validity is weaker here; elements of the definition (like indigenization pressure) appear to bring in elements other than the strict repudiation of contracts, and the enumerated list of reasons for contract repudiation is quite thin – additional unstated reasons for contract repudiation could include fiscal crisis, war or simply the whims of an unaccountable leader.

Ultimately, Knack's analysis is undermined by insufficient attention to theory at the outset. He defines his systematized concept poorly, and skips directly from construct to operationalization. In addition, the data are compromised by a subjective (and proprietary) measurement strategy.

We will now consider criterion-related validity. Criterion validity is typically assessed via correlation of the measures in question with other well-established measures and deducing that the measure is associated with other variables in a theoretically predictive manner (Trochim and Donnelly 2007). This will involve assessment of the IRIS dataset's predictive validity, concurrent validity, convergent validity and discriminant validity. As we have said, the dataset purports to measure quality of governance. Criterion validity is most useful in a context with strong theory and established measures, which are largely (though not entirely) absent.

If the dataset exhibits high predictive validity, it should be able to predict whether a country is well-governed or not. This appears to be the case – countries scoring highest and lowest on all 6 measures are indeed the usual suspects, ranging from Finland and Singapore at one end to Bangladesh and Haiti at the other. The data also should track with other quality-of-governance indices. A crude analysis reveals that this is indeed the case. The dataset is highly correlated with Freedom House's *Freedom in the World* data, which are a comparative assessment of political rights.

Strong concurrent validity would imply that the dataset should be able to distinguish between well-governed and poorly-governed countries, measured according to "gold standards". Corruption tracks Transparency International's widely-accepted (though itself problematic) measure, with a few surprises that we lack the space to fully explore. Other measures such as bureaucratic efficiency and expropriation risk seem to lack a widely-accepted gold standard. Quality of governance should also correlate with economic growth, if prevailing theories are correct (e.g. Kaufmann and Kraay 2008).

If the IRIS dataset exhibits strong convergent validity, this would imply high average interitem correlation among the individual measures, as well as correlation across individual years within the dataset. Both of these appear to be the case, with a surprising exception. The measure of ethnic tensions (added after Knack's 1995 paper) appears to be significantly less correlated with the other five than each of those five measures are with each other, raising questions about the decision to include it.

Correlation

	Corruption	Rule of Law	Bureaucratic Quality	Ethnic Tensions	Contract Repudiation	Expropriation Risk
Corruption	1.0000	0.7157	0.7655	0.4702	0.6080	0.5877
Rule of Law		1.0000	0.7550	0.6051	0.7574	0.7863
Bureaucratic Quality			1.0000	0.4283	0.7153	0.6662
Ethnic Tensions				1.0000	0.5473	0.5313
Contract Repudiation					1.0000	0.8773
Expropriation Risk						1.0000

It is difficult to assess the discriminant validity of the IRIS dataset. Discriminant validity predicts that measures of different constructs will not correlate. However, institutional quality is deeply involved in every aspect of governance and in many quantifiable social and economic indicators, making a true comparison problematic.

Reliability is also difficult to assess for these data because we would need to know more about PRS's procedures for scoring. For example, we are unable to measure inter-observer reliability or parallel forms reliability. In addition, it is impossible to repeat the experiment to get new data – we (fortunately) can't re-run the 1980s. However, the internal consistency of each variable over time demonstrates repeatability in the measures. Covariance analysis suggests relatively robust reliability, with an important exception for the Ethnic Tensions measure.

Covariance

	Corruption	Rule of Law	Bureaucratic Quality	Ethnic Tensions	Contract Repudiation	Expropriation Risk
Corruption	2.1424	1.7291	1.7483	1.1063	2.0412	1.9442
Rule of Law		2.7241	1.9446	1.6055	2.8943	2.9610
Bureaucratic Quality			2.4349	1.0745	2.5644	2.3536
Ethnic Tensions				2.5843	2.0500	1.9608
Contract Repudiation					5.4876	4.7443
Expropriation Risk						5.3287

An important overall note concerning the reliability of these operationalizations is that they are all subjective scores rather than objective measures. As a result, the noise and bias components of the true score will be relatively large. The scoring process can be expected to introduce noise, but its extremely subjective nature is also likely to introduce bias. By translating their intuitions into subjective scores, the staff of ICRG have made their data much more amenable to statistical analysis, but there is significant risk that this process will introduce sufficient bias to undermine the utility of the measures.

References

Knack, S. and Keefer, P. (1995) Institutions and Economic Performance: Cross-Country Tests Using Alternative Institutional Measures. Economics & Politics, 7, 207-227. https://doi.org/10.1111/j.1468-0343.1995.tb00111.x
Trochim, B. and Donnelly, J. (2007). *The Research Methods Knowledge Base*. Atomic Dog Publishing, 2007.
Davis, K. and Fisher, A. (eds.) (2012). *Governance by Indicators: Global Power through Quantification and Rankings*. Oxford University Press 2012.
The PRS Group (1993). *IRIS Dataset (IRIS-3)*. Compiled by Knack, S. and Keefer, P. https://epub.prsgroup.com/products/icrg/iris-dataset
The PRS Group (2012). "International Country Risk Guide Methodology", https://www.prsgroup.com/wp-content/uploads/2012/11/icrgmethodology.pdf
The PRS Group (2014). *Guide to Data Variables*. https://epub.prsgroup.com/list-of-all-variable-definitions
International Country Risk Guide (ICRG) Researchers, 2013, "International Country Risk Guide (ICRG) Researchers Dataset", https://hdl.handle.net/1902.1/21446, Harvard Dataverse V3. Kaufmann, D. and Kraay, A. (2008)
Governance Indicators: Where Are We, Where Should We Be Going? The World Bank Research Observer, 23, 1-30. https://doi.org/10.1093/wbro/lkm012
Glaeser, E.L., La Porta, R., Lopez-de-Silanes, F. and Shleifer, A. (2004) Do Institutions Cause Growth? Journal of Economic Growth, 9, 271-303. https://doi.org/10.1023/B:JOEG.0000038933.16398.ed