10-1-2011

Through the Looking Glass: Understanding Social Science Norms for Analyzing International Investment Law

Susan Franck
American University Washington College of Law, sfranck@wcl.american.edu

Calvin Garbin
University of Nebraska at Lincoln

Jenna Perkins
University of Nebraska at Lincoln

THROUGH THE LOOKING GLASS: UNDERSTANDING SOCIAL SCIENCE NORMS FOR ANALYZING INTERNATIONAL INVESTMENT LAW

Susan D. Franck,[*] *Calvin P. Garbin*[**] *and Jenna M. Perkins*[***†]

> "We've learned from experience that the truth will come out. Other experimenters will repeat your experiment and find out whether you were wrong or right. Nature's phenomena will agree or they'll disagree with your theory. And, although you may gain some temporary fame and excitement, you will not gain a good reputation as a scientist if you haven't tried to be very careful in this kind of work."[1] – Richard P. Feynman, Nobel Prize Winner in Physics (1965)

Social science promotes the acquisition of knowledge based upon data that we derive from observable and knowable phenomena. When social science methods are being employed in a new context—such as the assessment of international investment law—there is value in exploring the underlying assumptions and normative baselines of the overall enterprise. We therefore welcome the dialogue generated by Professor Van Harten in his *Yearbook* contribution, *The Use of Quantitative Methods to Examine Possible Bias in Investment Arbitration* ("*Yearbook Contribution*"), as it echoes the areas for future consideration identified in 2008.[2] We wish to offer a framework for productive academic discourse and social science insights in order to promote thoughtful future research and commentary in international investment law. Clarifying the social science framework is necessary for constructive and thoughtful assessment of empirical research on international investment law. While we recognize this approach makes *Through the Looking Glass* somewhat unusual, our hope is that we can promote the evolution of knowledge by situating the discussion within a larger literature.

The *Yearbook Contribution* raises various issues about a 2009 *Harvard International Law Journal* article.[3] We begin by identifying our core concerns with the *Yearbook Contribution*, namely that it does not identify that: (1) the *Harvard International* article stated the limitation of the OECD-measure and, all analyses—even analyses using the measure that Van Harten suggests—found no relationship between development variables and outcome;[4] (2) small-

[*] Associate Professor of Law, Washington & Lee University School of Law.

[**] Professor of Psychology, University of Nebraska-Lincoln.

[***] J.D. University of Nebraska-Lincoln College of Law (2010). Ph.D. in Law and Psychology Candidate, University of Nebraska-Lincoln.

[†] The authors are grateful to Eve M. Brank, Miriam A. Cherry, Joni Hersh, David S. Law, Tonya L. Putnam, Beth A. Simmons and Sandra Wiebe for comments on this article. We also express our appreciation to the Washington & Lee University of Law Library Staff for their unfailing research support.

[1] RICHARD P. FEYNMAN & RALPH LEIGHTON, "SURELY YOU'RE JOKING, MR. FEYNMAN!": ADVENTURES OF A CURIOUS CHARACTER 312 (1986).

[2] Susan D. Franck, *Empiricism and International Law: Insights for Investment Treaty Dispute Resolution*, 48 VA. J. INT'L L. 767, 774, 784, 787-90, 793, 794 (2008).

[3] Susan D. Franck, *Development and Outcomes of Investment Treaty Arbitration*, 50 HARV. INT'L L.J. 435 (2009) [hereinafter Franck, *Harvard International* article].

[4] *See infra* Section II and page 7.

n studies of international economic law phenomenon, including investment treaty arbitration, are valuable;[5] (3) the *Harvard International* article acknowledged—in more than thirty separate instances—the limitations of the data and the need for replication;[6] (4) while we agree that perceived bias may be a constructive area for future research, the *Harvard International* article sought to first identify observable bias;[7] and (5) the results were population parameters for the pre-2007 dataset, and the *Harvard International* article's inferences beyond that population were neither unreasonable nor unfounded.[8]

In an effort to take methodological concerns seriously and develop a common framework for future discussions, we address the issues raised above in four parts. First, we describe the value of social science in international investment law. Second, we explore the critique of one variable used in the *Harvard International* article and find, upon replicating the analyses using Van Harten's proposed definition of development, there was still no reliable relationship between development status and outcome. Yet, recognizing that there can be different ways to measure development, we suggest refinement of the existing variables for future analyses. Third, we identify social science norms related to population and sample size, particularly for parameters on the analyzed population and other inferences for ruling out the presence of statistically large and medium sized effects. While the dataset of investment arbitration awards provided relatively low statistical power and much additional data would be required to detect the smallest possible effects, it is likely to take several decades (perhaps in the order of 50 years) before the necessary data exists to make such analysis possible. As a normative matter, the authors believe that is too long to wait before conducting analyses on an area with critical implications for private and public international law. The better course is to conduct analyses that are methodologically sound ex ante, report the analyses and acknowledge their limitations so that future researchers can consider and expand upon the baseline. Finally, we explore opportunities for the evolution of social science research of international investment law. We conclude that the value of both careful research and informed critiques is prudent in an area of international economic activity.

I. FUNDAMENTALS IN SOCIAL SCIENCE RESEARCH

The objective of using social science methodologies to study international law phenomenon is to promote the systematic gathering and analysis of data to test assumptions about reality in an effort to make the best—and most informed—normative choices. Human beings gather, analyze

[5] *See infra* Section III(A); *see also id.* at 4-5.
[6] *See infra* Appendix I.
[7] *See infra* pages 22, 24.
[8] *See infra* Section III(B), III(C).

and act on information in ways that can be a by-product of cognitive biases[9] and lead to inadvertent errors.[10] Recognizing this risk of error and in the hope of improving decisions, social science offers empirical tools to correct our potential misperceptions and to promote thoughtful and informed decisions. Although empirical analysis need not focus on a single methodology, as the *Yearbook Contribution* focuses upon quantitative research, we consider the value of quantitative analysis of investment treaty disputes. In an effort to do this in a methodologically rigorous manner and to create a common framework of social science norms, this Section has three goals. First, it places social science research into a broader context about the search for knowledge to aid normative choices. Second, it argues that quantitative methodologies are valuable. Third, it explores key social science concepts.

A. The Contextual Value of Empirical Knowledge

Scholarship is about the search for knowledge in an effort to understand, to act upon and to interact with our world. There are, however, different ways to gain knowledge. Knowledge can be derived from: (1) instinct or intuition, (2) reliance on an authority, (3) the use of logic or rationalism, or (4) use of empirical science.[11]

Instinct or intuition is an impressionistic way of amassing knowledge. It can have a powerful effect on our beliefs,[12] but yet these beliefs can be wrong given our susceptibility to cognitive biases.[13] Authority involves gaining knowledge by relying on information from a respected source, and believing that information to be true.[14] Even if authorities express their beliefs forcefully, there are challenges to knowledge evolution if the beliefs of authorities are wrong. Logic is third way of amassing knowledge. Although it

---

[9] Cognitive biases that can affect our data collection, analysis, and decision-making include: (1) confirmation bias; (2) expectation bias; (3) selective perception; (4) the projection bias; and (5) blind spot bias. *See, e.g.*, DAN ARIELY, PREDICTABLY IRRATIONAL (2008); JONATHAN BARON, THINKING AND DECIDING 171–77, 191–92, 205-07, 221–27 (4th ed. 2008); SCOTT PLOUS, THE PSYCHOLOGY OF JUDGMENT AND DECISION MAKING (1993); Ronald Chen & Jon Hanson, *Categorically Biased: The Influence of Knowledge Structures on Law and Legal Theory*, 77 S. CAL. L. REV. 1103 (2004).

[10] *See generally* ARIELY, *supra* note 9; Dan M. Kahan, *Two Conceptions of Emotion in Risk Regulation*, 156 U. PENN. L. REV. 741 (2008).

[11] DONALD H. MCBURNEY & THERESA L. WHITE, RESEARCH METHODS 2-3 (8th ed. 2010); *see also* SHERRI L. JACKSON, RESEARCH METHODS AND STATISTICS: A CRITICAL THINKING APPROACH 6-7 (3d ed. 2009); ANTHONY M. GRAZIANO & MICHAEL L. RAULIN, RESEARCH METHODS: A PROCESS OF ENQUIRY 7-11 (7th ed. 2010).

[12] McBurney & White, *supra* note 11, at 3-4.

[13] *See* KEITH E. STANOVICH, HOW TO THINK STRAIGHT ABOUT PSYCHOLOGY 13 (9th ed. 2010).

[14] McBurney & White, *supra* note 11, at 2-3; Jackson, *supra* note 11, at 7.

helps us draw inferences and understand when conclusions are improper, logic is limited when the factual predicate—namely the intellectual premise—for the reasoning is incorrect.[15] Logic and rational analysis, therefore, can benefit from insights from a process that permits the falsification of underlying factual predicates; this, in turn, promotes a more informed (but not perfect) basis for related analyses, conclusions and normative choices.

The scientific method attempts to gain knowledge from reality-based observation that defines a problem, forms a hypothesis, collects data, draws inferences and communicates the findings.[16] Where there are clear, sound and systematic ex ante research protocols, the knowledge derived from that process is normatively preferable to chance alone and more verifiable than intuition or authority. What makes science unique is a "willingness to change [initial] beliefs based on objectively obtained empirical evidence derived from their method of enquiry."[17] The real value of empiricism is the fundamental freedom to admit errors. The intensity of a belief in truth (or falsity) of a hypothesis about is irrelevant.[18] Rather, the scientific enquiry is about "'making mistakes in public—making mistakes for all to see, in the hopes of getting others to help with the corrections' . . . and continually adjusting theory when data do not accord with it."[19] Fundamentally, a social scientific approach to the analysis of legal phenomena means that we can attempt to disprove falsehoods and can adjust to reality as reality adjusts around us.

B. A Normative Preference for Empirical Perspectives of Investment Treaty Disputes

International investment law faces a fundamental question: How do we choose to amass knowledge about an area of fundamental economic activity in a time of global economic transition? We believe that reliance on scientific methodologies with sound ex ante research protocols (whether the research is quantitative, qualitative, or mixed-methods), is worth the resulting costs[20] and preferable to instinct, intuition or chance alone.[21]

Reliance on quantitative analyses utilizing aggregate data may be challenging for lawyers, particularly as many lawyers are trained to focus on individual cases and to identify definitive truth. The essence of quantitative

---

[15] McBurney & White, *supra* note 11, at 3; Jackson, *supra* note 11, at 8-9.

[16] McBurney & White, *supra* note 11, at 5-9; Stanovich, *supra* note 13, at 9-10; ROBERT M. LAWLESS, JENNIFER K. ROBBENNOLT & THOMAS S. ULEN, EMPIRICAL METHODS IN LAW 7-21 (2010).

[17] McBurney & White, *supra* note 11, at 10.

[18] PETER B. MEDAWAR, ADVICE TO A YOUNG SCIENTIST 39 (1979).

[19] Stanovich, *supra* note 13, at 28 *quoting* DANIEL C. DENNETT, DARWIN'S DANGEROUS IDEA: EVOLUTION AND THE MEANING OF LIFE 380 (1995).

[20] *See, e.g.,* Franck, *Empiricism and International Law*, *supra* note 2, at 774.

[21] ROBERT S. PINDYCK & DANIEL L. RUBINFELD, ECONOMETRIC MODELS & ECONOMIC FORECASTS xvi-xv (4th ed. 1998).

social science norms is that, in an effort to get better information, one must avoid the temptation to focus on isolated and unrepresentative examples and recognize there will not be quick, easy, definitive answers. Rather, "an empirical approach is one of ongoing inquiry. Each new study makes an incremental factual advance, building on the empirical work that has come before it and raising new questions for future inquiry".[22] No single piece of research is perfect or irrefutable for all time.

Social science rejects "truthiness"[23] in favor of data—and acknowledges that data and ideas can (and arguably must) evolve over time. Some of the most fundamental breakthroughs in society have come from using knowledge that is derived from systematic, empirical observation to re-assess conventional wisdom.[24] Recognizing the inevitable challenges but understanding the corresponding benefits, it is necessary to conduct empirical analyses to assess the existing baselines and use what we know to consider and evaluate where we wish to be normatively in the future. Particularly in the context of investment treaty arbitration with a small but growing population of disputes, it is constructive to take snapshots over time, to consider what we know and to consider the implications for the future as we add to our knowledge over time.

C. Defining Fundamental Terms

The *Yearbook Contribution* uses the terms "valid," "reliable" and "transparent." In light of social science norms, we prefer to begin with a clear statement[25] of how we understand these core terms. As studying investment treaty dispute resolution is relevant[26] to policy debates, the *Harvard International* article was an effort to take seriously the concerns of NGO and government officials. The objective was to use data, measures and models to

---

[22] Lawless, et al., *supra* note 16, at 15.

[23] Stanovich, *supra* note 13, at 29 *quoting* FARHAD MANJOO, TRUE ENOUGH: LEARNING TO LIVE IN A POST-FACT SOCIETY 198 (2008) (emphasis in original) (""Stephen Colbert coined the term 'truthiness.' Truthiness is the 'quality of a thing *feeling* true without any evidence suggesting it actually was'").

[24] For example, cognitive psychologists who identified flaws in eyewitness testimony and created research that, in combination with DNA testing, has permitted wrongly convicted individuals to go free. *See, e.g.,* Gary L. Wells & Deah S. Quinlivan, *Suggestive Identification Procedures and the Supreme Court's Reliability Test in Light of Eyewitness Science: 30 Years Later*, 33 LAW. & HUM. BEHAV. 1 (2009); Gary L. Wells, *Theory, Logic, and Data: Paths to a More Coherent Eyewitness Science*, 22 APPLIED COGNITIVE PSYCH. 853 (2008).

[25] *See* Lee Epstein & Gary King, *The Rules of Inference*, 69 U. CHI. L. REV. 1, 74 (2002) ("This process of "clarification" is sometimes called "operationalizing," "operationally defining," or more simply, "defining" the concepts."); *see also* Graziano & Raulin, *supra* note 11, at 74-75, 152; Jackson, *supra* note 11, at 57-58.

[26] Lawless, et al., *supra* note 16, at 27-28.

aid in the evaluation of their hypothesis—namely that there was bias against the developing world—because, if true, it would be a cause for concern.

1. Validity

As a general matter, a measure is considered valid if it measures what it purports to measure.[27] Yet there are different types of validity, including but not limited to internal validity, external validity and construct validity.

Internal validity is concerned with the procedural control a researcher has and maintains over his/her study, typically whether internal research procedures have inadvertently introduced a confounding variable or other error.[28] In assessing internal validity, it is important to examine whether factors other than the variable of interest could be responsible for the results.[29]

External validity refers to the generalizability of results to "other participants, settings, and times."[30] A study is generalizable when it has relevance to real life.[31] The use of archival data—like data in investment arbitration awards—is valuable because it derives from the real world activities. Although the *Yearbook Contribution* incorrectly labels the concerns as issues of "internal validity",[32] the real concern seems to be with questions about the generalizability of the results. External validity concerns about generalizability of pre-2007 data to the current population are reasonable. For the then-known population of public awards analyzed, the results were population parameters. Concern for inferences about the larger population (namely private awards, post June 1, 2006 awards or future awards) is why the *Harvard International* article called for caution and replication.[33]

Construct validity is the "degree to which a study measures and manipulates the underlying psychological elements that the researcher claims to be measuring and manipulating."[34] In assessing construct validity, it is

---

[27] Phoebe C. Ellsworth & Richard Gonzales, *Questions and Comparisons: Methods of Research in Social Psychology* 35, *in* THE SAGE HANDBOOK OF SOCIAL PSYCHOLOGY 35 (Michael A. Hogg & Joel Cooper eds., 2007).

[28] Graziano & Raulin, *supra* note 11, at 164; Jackson, *supra* note 11, at 207-12; Lawless, et al., *supra* note 16, at 36-37.

[29] Kathryn C. Oleson & Robert M. Arkin, *Reviewing and Evaluating a Research Article* 66, *in* The PSYCHOLOGY RESEARCH HANDBOOK: A GUIDE FOR GRADUATE STUDENTS AND RESEARCH ASSISTANTS 66 (Frederick T. L. Leong & James T. Austin eds., 2d ed., 2006).

[30] SAVILLE, *supra* note 2, at 90; Lawless, et al., *supra* note 16, at 39-40.

[31] SAVILLE, *supra* note 2, at 90. This is also sometimes referred to as ecological validity. Lawless, et al., *supra* note 16, at 39.

[32] The *Yearbook Contribution* states, "This limitation arose from a lack of data. It affects the internal validity of the study by removing the prospect of results that could support inferences about the expected connections between development status and outcome." Van Harten, YB Contribution, at 16. Internal validity relates to the integrity of the internal research protocols—not a lack of data.

[33] *See* Appendix I.

[34] SAVILLE, *supra* note 2, at 85; *see also* Epstein & King, *supra* note 25, at 87-88;

important to determine whether the researcher operationalized his/her variables properly.[35] Has the researcher clearly explained how the variable was defined? Were alternatives considered and tested? Sometimes, it is difficult to operationalize a variable of interest—such as what it means to be part of the developed or developing world—and a proxy variable is substituted.[36] Put another way, "different researchers may choose to define similar constructs in different ways."[37] Scholars may disagree about the relative merits of certain constructs, but if there are multiple variables attempting to assess a construct and those definitions are stated expressly, that is a step in the right direction.

Concern for construct validity is appropriate. The *Yearbook Contribution* raises two issues about the *Harvard International* article, namely the use of OECD membership as a proxy for "development status" and adjusting amounts awarded for inflation. We explore the OECD variable in Section II. In all models analyzed, including those using Van Harten's construct, the definition did not affect the results—the data did not reveal a significant relationship between development status of respondents, development status of presiding arbitrators and outcomes.[38] As regards the inflation adjustment, replicating the models in the *Harvard International* to adjust for inflation, the results still did not reveal a significant relationship between development-related variables and amounts awarded. The *Harvard International* article coded the raw data from the actual date of an award.[39] For present purposes, we used two separate indices, namely the Consumer Price Index and Gross Domestic Product per capita, to adjust all awards to the common year of 2006 (i.e. the close of the dataset). Appendix II describes the sixteen different models we analyzed to consider whether adjusting for inflation made a difference in the findings.[40] All of those analyses replicated

---

Lawless, et al., *supra* note 16, at 40-41.

[35] Kathryn C. Oleson & Robert M. Arkin, *Reviewing and Evaluating a Research Article*, *in* The PSYCHOLOGY RESEARCH HANDBOOK: A GUIDE FOR GRADUATE STUDENTS AND RESEARCH ASSISTANTS 59, 67 (Frederick T. L. Leong & James T. Austin eds., 2d ed., 2006); Stanovich, *supra* note 13, at 39.

[36] BIOLOGICAL PSYCHIATRY 18 (H. A. H. D'haenen, John A. den Boer & Paul Wilner, eds.).

[37] Lawless, et al., *supra* note 16, at 51.

[38] *See infra* notes 65-70.

[39] The *Harvard International* article stated that it relied upon the archival dataset from the *North Carolina* article. Franck, *Harvard International* article, *supra* note 3, 454 at n.105. Although the *Yearbook Contribution* mentions that article only once, Van Harten, YB Contribution, page 11 at n.24, the *North Carolina* article explained, "The author later used a single Web site to convert the foreign currencies into a U.S. dollar amount (using the date of the award as the relevant conversion date) to create a common currency." Susan D. Franck, *Empirically Evaluating Claims About Investment Treaty Arbitration*, 86 N.C. L. REV. 1, 22 n.98 (2007) [hereinafter Franck, *Evaluating Claims*].

[40] Appendix II also addresses the concerns raised by the *Yearbook Contribution* in

the results in the *Harvard International* article, namely that there was no relationship between development variables and amounts awarded. Even with the limitation of low power, this demonstrates that the results in the *Harvard International* article were robust.

2. Reliability

Reliability relates to the consistency of a measurement instrument.[41] "Reliability is the extent to which it is possible to replicate a measurement, reproducing the same value (regardless of whether it is the right one) on the same standard for the same subject at the same time."[42] Measures are reliable if they produce the same value over time given the same inputs.[43] For example, if you use a scale to measure the weight of a rock, the measure—namely, the scale—is reliable if the same rock weighs the same amount each time it is placed on a scale.[44] Producing consistent results over time is not only reliability, however. It is also a function of *replicability*.

3. Replicability and Transparency

Transparency—meaning the identification of research protocols and choices—is a means to the end of replicability.[45] The "confirmation of research findings through replication by other researchers is an essential part of scientific methodology."[46] Researchers try to use identical or similar protocols to collect and analyze data; and the objective is usually to replicate

---

connection with winsorizing data. The skewness in the winzorized data was less than raw data, trimmed data or even squared data, which warranted its use in the *Harvard International* article. When replicating the research by analyzing both log and inverse transformed data with skewness levels of less than 1.0, the results were the same. There was no reliable relationship between the variables of interest and amounts awarded. Even with low power, the congruity in results demonstrates the results in the *Harvard International* article were robust.

[41] Stanovich, *supra* note 13, 38; Graziano & Raulin, *supra* note 11, at 78-79; *see also* Jackson, *supra* note 11, at 65.

[42] Epstein & King, *supra* note 25, at 83; *see also* Lawless, et al., *supra* note 16, at 42.

[43] *See* Epstein & King, *supra* note 25, at 85 ("The key to producing reliable measures is to write down a set of very precise rules for the coders . . . to follow—with as little as possible left to interpretation and human judgment."); *see also* Franck, *Evaluating Claims*, *supra* note 39, at 23 (describing the coding process).

[44] The reliability of the OECD measure over time is a constructive point. Theoretically, OECD membership—or World Bank development classification—may shift over time as states join the OECD and/or their gross national income (GNI) increases. One would hope, states would evolve over time and poverty might be eliminated or minimized. It would be unfortunate if the *Yearbook contribution*'s search for "reliability" means that we do not take into account the idea that the developing world may indeed develop.

[45] *See* Epstein & King, *supra* note 25, at 38 (2002); Stanovich, *supra* note 13, at 10.

[46] William G. Dewald, Jerry G. Thursby & Richard G. Anderson, *Replication in Empirical Economics: The* Journal of Money, Credit and Banking *Project*, 76 AM. ECON. REV. 587, 587 (1986); *see also* Lawless, et al., *supra* note 16, at 26.

aspects of the original study while expanding the research to add something new. A norm exists by which researchers can ask to use, and add to, data collected by others, either directly from the authors or from a journal that has published work, subject to concerns about proprietary nature of the data.[47] The dataset for the *Harvard International* article has been publicly available from the *North Carolina Law Review* since 2007. [48]

Transparency typically involves providing enough information to replicate the research in some manner or to understand the implications of a study, including internal validity, external validity, construct validity and the like.[49] Providing information aids the assessment of the underling research as well as the value of inferences drawn from the data, methods and analysis. Nevertheless, the *Yearbook Contribution* objects to certain results being "reported only in a series of footnotes or elliptically in the body of the article."[50] We recognize that: (1) consumers of scholarship are capable of reading all the words on a page, (2) standards of the American Psychological Association—let alone conventions in U.S. law reviews—do not require everything to be in the text and permit (and sometimes require) footnotes, and (3) the *Harvard International* article placed key information in the text and used footnotes to add complementary information or clarify details.

The *Harvard International* article is wholly transparent about the time-bounded nature of its underlying archival data. The article explained its measurement system, described the statistical choices and offered results. Further, it explained the limitations and recommended replication to assess whether the results would continue to hold over time. Given the nature of the data and the statistical power of the research, as the population grows over time, it is entirely possible that the results of the research may change and the *Harvard International* article may only be an initial historical snapshot. Future analyses may also find different or complementary nuances as we develop more sophisticated measures and models in the quest to minimize statistical error. Productive areas of future scholarship on investment law may focus on such historical shifts and methodological innovations. In the interim, we are all free to—based upon the data and results—draw independent inferences and make our own normative choices. It is the purview of

---

[47] AMERICAN PSYCHOLOGICAL ASSOCIATION, PUBLICATION MANUAL 12-13 (6th ed. 2010) [hereinafter APA Manual].

[48] Franck submitted her data to the *North Carolina Law Review* and made it available upon request. Franck, *Evaluating Claims*, *supra* note 39, at 17, 20 (2007). Van Harten received the relevant subset of the data upon request and has acknowledged this. Van Harten, YB Contribution 33, n.71. The proprietary dataset is also now available online in connection with the publication of *Rationalizing Costs in Investment Treaty Arbitration*.

[49] *See, e.g.,* MARK R. LEARY, INTRODUCTION TO BEHAVIORAL RESEARCH METHODS 365-70 (3d ed., 2001).

[50] Van Harten, YB Contribution at 18.

academics to debate those inferences, and it is the occupation of policy makers to make choices. Even if we recognize that research will never capture the complexity of reality,[51] it would be irresponsible to ignore empirical evidence generated in accordance with scientific best practices simply because evidence does not comport with conventional wisdom or even a research hypothesis.

II. MEASURING DEVELOPMENT

One of the most challenging aspects of any social science inquiry is construct validity, namely finding an indicator to capture a social phenomenon in which a researcher is interested. "It may be the case that the variable being explained cannot be measured accurately, either because of data collection difficulties or because it is inherently unmeasureable and a proxy variable must be used in its stead."[52] The meaning of "Development Status" is instructive. There is not a consistent legal definition of this concept; and development means different things to different people. For example, the World Trade Organization does not define development; rather, it permits member states to self-define their development status.[53] The lack of a consistent definition also creates confusion in international environmental law.[54] When there is no predefined and exclusive measure for legal phenomenon, it is preferable to use definitions that come from "judgments made for entirely different purposes by *other researchers*."[55] The *Harvard International* article does this with the OECD variable. Hallward-Driemer's article[56] uses OECD membership to analyze development. As the variable

---

[51] *See* Franck, *Empiricism and International Law*, *supra* note 2, at 790.

[52] PETER KENNEDY, GUIDE TO ECONOMETRICS 5 (5th ed. 2003); *see also* Joshua B. Fischman & David S. Law, *What Is Judicial Ideology, and How Should We Measure It?*, 29 WASH. U. J.L. & POL'Y 133, 143-45 (2009).

[53] *See* World Trade Organization, Who are the developing countries in the WTO?, at http://www.wto.org/english/tratop_e/devel_e/d1who_e.htm ("There are no WTO definitions of 'developed' and 'developing' countries. Members announce for themselves whether they are 'developed' and 'developing' countries"); Anu Bradford & Eric A. Posner, *Universal Exceptionalism in International Law*, 52 HARV. INT'L L.J. 1, 32 n.159 (2011); Andrew D. Mitchell & Joanne Wallis, *Pacific Pause: The Rhetroic of Special & Differential Treatment, the Reality of WTO Accession*, 27 WISC. INT'L L.J. 663, 696-97 (2010).

[54] Benjamin L. Liebman, *Autonomy Through Separation?: Environment Law and the Basic Law of Hong Kong*, 39 HARV. INT'L L.J. 231, 261-62 (1998) ("Environmental agreements are inconsistent in their definitions of development status . . . various treaties use different standards to define developed and developing countries.") (footnotes omitted).

[55] GARY KING, ROBERT O. KEOHANE, SIDNEY VERBA, DESIGNING SOCIAL ENQUIRY: SCIENTIFIC INFERENCE IN QUALITATIVE RESEARCH 157 (1994) (emphasis in original).

[56] *See* Mary Hallward-Driemeier, *World Bank, Do Bilateral Investment Treaties Attract FDI? Only a Bit…and They Could Bite* (2003), http://econ.worldbank.org/files/29143_wps3121.pdf at 8, 9, 11-12, 13 (referring to "rich OECD countries", distinguishing between OECD members and "developing countries", analyzing FDI flows "from OECD countries to developing country hosts"); *see also* Jeswald W. Salacuse & Nicholas P. Sullivan, *Do BITs Really Work?: An Evaluation of Bilateral*

derives from a State's decision to ratify the OECD Convention and other social scientists studying investment employ the classification, it was appropriate to use OECD membership to evaluate development.

The *Harvard International* article described the countries that were OECD Convention signatories in 2006.[57] It also included an alternative variable to measure development status—namely the World Bank's classification based upon Gross National Income (GNI)—as a cross-check.[58] The *Harvard International* article explained: "Both OECD status and World Bank status were assessed in order to address different conceptions of what development can mean . . . For example, Mexico is a member of the OECD but is not classified as a High Income country."[59] The *Yearbook Contribution*'s claim that the limitations of the OECD variable "were not disclosed"[60] is therefore incorrect. Of course, the definition of development, whether based on OECD membership or other indicators, is an area for dialogue. Yet, the *Yearbook Contribution* suggests a new way of measuring "development" that is not based upon any existing definition or research standards known to the authors. Although we welcome contributions that better capture the construct of development, such a variable should satisfy the criteria elaborated below.

First, the variable would have to specify its groupings of states.[61] The *Yearbook Contribution* refers to "East Bloc" countries without defining the states. Second, the variable should be backed by a reasoned discussion explaining what makes "Mexico, South Korea, Turkey, and [the] former East Bloc," different from other OECD members. Presumably this would be a function of GNI or some other factor related to economic resources, yet it would be helpful to conform to social science practice of explaining why there

---

Investment Treaties and Their Grand Bargain, 46 HARV. INT'L L.J 67, 104-15 (2005); UNITED NATIONS CONFERENCE ON TRADE AND DEVELOPMENT, BILATERAL INVESTMENT TREATIES IN THE MID-1990S, 103 (1998); Eric Neumayer & Laura Spess, *Do Bilateral Investment Treaties Increase Foreign Direct Investment in Developing Countries?*, 17 (May 2005), *at* http://129.3.20.41/eps/if/papers/0411/0411004.pdf.

[57] In 2006 and until 2010, the OECD members were Australia, Austria, Belgium, Canada, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, South Korea, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, Slovak Republic, Spain, Sweden, Switzerland, Turkey, United Kingdom, and United States. OECD, List of OECD Member countries - Ratification of the Convention on the OECD, *at* http://www.oecd.org/document/58/0,2340,en_2649_201185_1889402_1_1_1_1,00.html. Franck used a State's decision to join the OECD as to classify a state. Franck, *Harvard International* article, *supra* note 3, at 455 n.110; *see also* Franck, *Evaluating Claims*, *supra* note 39, at 28 n.125.

[58] Franck, *Harvard International* article, *supra* note 3, at 455 n.112.

[59] *Id.,* 455 n.109.

[60] Van Harten, YB Contribution at 33; *see also id.* at 15.

[61] *See supra* note 25 (discussing the operationalization of variables).

is "fidelity between the construct and the measure".[62] Third, the variable should indicate the scale or categories proposed and why those are normatively and statistically valid constructs.[63] Finally, the variable should clarify how it is meaningfully different—and of greater value—than the more granular variable of World Bank Status. Understanding the nature of the variable and the value it seeks create would be a welcome addition to the larger dialogue on the meaning of "development." In other words, new variables should establish "facial validity, unbiasedness, and efficiency."[64]

Presuming that Van Harten wished to create his own variable that excludes OECD states from the OECD measure, for the purpose of this article, we created a "development status" variable that started with OECD members but excluded Mexico, the Czech Republic and the Slovak Republic.[65] We created this measure both for the respondent states and the presiding arbitrators.[66] However, even using that definition of "development status", there were no meaningful differences in outcomes, namely, at the macro level: (1) there was no reliable pattern of relationship between development status variables and winning or losing an arbitration, and (2) there was no reliable relationship between development status variables and amounts awarded.

1. Replicating the results on page 460 of the *Harvard International* article but using Van Harten's definition of "development status," there was still no pattern of relationship between development status and whether a respondent won or lost. When analyzing the set of cases with presiding arbitrators from non-developed countries, there was no pattern of relationship between developed or developing states ($\chi^2(1) = 0.049$; $p = 0.83$; $r = 0.06$; n = 14). In other words, in cases with presiding arbitrators from developing countries, it was not possible to conclude that the

[62] Lawless, et al., *supra* note 16, at 41.

[63] The *Yearbook Contribution* refers to various categories including developed, developing and transition economies. Although the authors asked for clarification of the definition of measurement for respondent states, none has been forthcoming.

[64] Epstein & King, *supra* note 25, at 89.

[65] The *Yearbook Contribution* focuses on "Mexico, South Korea, Turkey, and the former East Bloc countries". Only three OECD members (Mexico, Czech Republic and Slovak Republic) had awards in the dataset. There were no awards against South Korea or Turkey.

[66] Classification of presiding arbitrators did not change as none were from Mexico, the Czech Republic or the Slovak Republic. There were two changes in cases against the Czech Republic: (1) *Lauder*, which involved a respondent win, and (2) *CME*, which involved a claimant win. There was one change to the Slovak Republic, namely, *CSOB*, which involved a respondent win on the treaty claim only. For Mexico, the state identified in the *Harvard International* article, eight cases were reclassified, namely: (1) *Azinian*, which involved a respondent win, (2) *Feldman*, which involved a claimant win, (3) *GAMI*, which involved a respondent win, (4) *Metalclad*, which involved a claimant win, (5) *Tecmed*, which involved a claimant win, (6) *Thunderbird*, which involved a respondent win, (7) *Waste Management I*, which involved a respondent win, and (8) *Waste Management II*, which involved a respondent win. For cases against Mexico, there were five government wins and three investor wins.

outcomes for developed and developing countries were meaningfully different. Similarly, when analyzing the set of cases with presiding arbitrators from developed (i.e. OECD) countries, there was likewise no relationship between states' development status and success or defeat in investment treaty arbitration ($\chi^2(2) = 0.118$; $p = 0.73$; $r = 0.06$; $n = 33$).[67] Again, the data could not confirm that tribunals with presiding arbitrators from the developed world decided cases against developed or developing states in a meaningfully different way. Using Cohen's conventions, the effect sizes for the models were less than small and the large *p*-values indicated the effects of the variables were far from statistically meaningful.

2. Replicating the results on page 465 of the *Harvard International* article but using Van Harten's definition of development, there was also no statistically significant mean difference in amounts awarded based upon development variables. First, using winzorized data, there was no interaction among the development status of the respondent, the development status of the presiding arbitrator and the amount awarded ($F(1,45) = 0.947$; $p = 0.37$; $r = 0.14$; $n = 49$).[68] Second, irrespective of

[67] Jacob Cohen was a statistician and quantitative psychologist whose scholarship explored power and effect size. Cohen's conventions—small = 0.10, medium = 0.30, large = 0.50—indicate the effect sizes of both analyses were less than small ($r = 0.06$). JACOB COHEN, STATISTICAL POWER ANALYSIS FOR THE BEHAVIORAL SCIENCES 25, 113-16, 124 (2d ed. 1988). If there is no statistically significant result in a hypothesis test and the effect size is < .10, generally it is unlikely the analysis suffers from a power problem (i.e. the sample is of an insufficient size to detect a significant effect) or the possible effect may be so small that it will be difficult to justify resources to research the issue. The authors recognize that, given the small effect size, the power of the analysis is likely less than .20, which means there is approximately an 80% chance of incorrectly retaining the null hypothesis. Similar to the analysis in the *Harvard International* article, this means that to reliably ascertain the presence of the effect, a sample of 1,562 awards would be needed. If the caseload grows at the arguable rate of 30 per year, achieving this size would take approximately 50 years.

[68] For the winzorized data, the mean awards were: (1) OECD Respondents with OECD Presiding Arbitrators = US$850,418 ($n = 5$); (2) OECD Respondents with Non-OECD Presiding Arbitrators = US$77,657 ($n = 2$); (3) Non-OECD Respondents with OECD Presiding Arbitrators = US$1,077,183 ($n = 30$); (4) Non-OECD Respondents with Non-OECD Presiding Arbitrators = US$2,177,070 ($n = 12$). For the raw data which includes statistical outliers and exhibited positive skewing, the mean awards were (1) OECD Respondents with OECD Presiding Arbitrators = US$850,418 ($n = 5$); (2) OECD Respondents with Non-OECD Presiding Arbitrators = US$77,657 ($n = 2$); (3) Non-OECD Respondents with OECD Presiding Arbitrators = US$10,334,852 ($n = 30$); (4) Non-OECD Respondents with Non-OECD Presiding Arbitrators = US$18,569,314 ($n = 12$). Although the descriptive raw data might raise a facial concern (namely what appear to be higher amounts awarded against non-developed states), the results indicated that the awards were statistically equivalent (namely all of the amounts awarded are not statistically different from zero). The difference fades when analyzing winzorized data. Although effect sizes suggest ruling out the possibility of either a

> whether the tribunal had a presiding arbitrator from a developed or developing country, there was no statistically significant effect on the amount awarded ($F(1,45) = 0.029$; $p = 0.87$; $r = 0.03$; $n = 49$). Third, irrespective of whether the respondent was a developed or a developing country, there was no statistically significant difference in the amount awarded against developed or developing countries ($F(1,45) = 1.462$; $p = 0.23$; $r = 0.17$; $n = 49$).[69] The same held true for raw amounts awarded.[70]

The replication demonstrates that even using Van Harten's proposed definition, the models could not identify a meaningful difference in outcome—whether as a function of who won or amounts awarded. Although two analyses exhibited low power that justifies replication before drawing definitive conclusions beyond the dataset studied, the results indicate that the findings of the *Harvard International* article were robust.

In 2010, there were changes to OECD membership. The complete list of new additions is: Chile, Estonia, Israel and Slovenia.[71] When conducting future research, it would be constructive to account for the expanded membership and to create a tighter nexus in the measurement of development classification. For example, as the date of the award is the operative date for many coding decisions, coding OECD and World Bank categories might use the date of the award rather than the cut-off date of the dataset. In this way, it would be possible to replicate and expand the analysis on the basis of expanded data and also consider whether there is any statistically meaningful difference between the different definitions of development status.

---

large or medium-sized effect of development on outcome, replication of research—with more data and better modeling is necessary and was recognized in the *Harvard International* article.

[69] For the interaction and the main effect of the development, the effect sizes were small to coming close to medium ($r = 0.14$; $r = 0.17$). This means that the statistical power was approximately .20 to .30, which indicates a 70-80% risk of committing a Type II error. In order to isolate the smaller of these effects, a sample of 686 arbitration awards is required. In contrast, for the main effect of the presiding arbitrator, the effect size was $r = 0.03$, which suggests that the failure to detect the miniscule effect was likely not a function of power.

[70] For the raw data using Van Harten's definition of "development status", there was no statistically significant interaction among development of respondent, development of the presiding arbitrator and outcome ($F(1,45) = 0.050$; $p = 0.082$; $r = 0.03$; $n = 49$. Likewise, there was neither a significant main effect for the development status of the presiding arbitrator ($F(1,45) = 0.035$; $p = 0.85$; $r = 0.02$; $n = 49$) nor a significant main effect for the development status of the respondent state ($F(1,45) = 0.485$; $p = 49$; $r = 0.10$; $n = 49$). With the miniscule effect sizes ($r < .10$), it is unlikely that the analyses suffer from a power problem; out of an abundance of caution, given that the power of the analysis is approximately .20 and there is a resulting 80% likelihood of a Type II error, replication is appropriate. The sample size needed to ascertain the small effect ($r = 0.10$), would be 1,562 awards.

[71] OECD, List of OECD Member countries – Ratification of the OECD Convention, http://www.oecd.org/document/58/0,2340,en_2649_201185_1889402_1_1_1_1,00.html (last visited Feb. 26, 2011).

III. SMALL POPULATIONS AND SAMPLE SIZE

Analysis of small datasets is not per se problematic. It is appropriate when handled with a proper understanding of data protocols. "Uncertainty and limited data should not cause us to abandon scientific research" [72] While there is inevitable uncertainty in social science, the possibility of statistical error does not mean we should reject or ignore research. Rather, the better approach is to consider the research along with its limitations. This Section first explains small-*n* research is appropriate for investment treaty arbitration. It then clarifies the results were parameters for the population studied and explores limitations of possible inferences (expressly identified in the *Harvard International* article). Even without being able to identify a reliable relationship, the analyses also revealed it is possible to reject the notion that development has a statistically large effect on outcome. Next, given the transparent methods and acknowledged limits, this Section then explains that the *Harvard International* article's inferences were reasonable.

*A. Small-*n *Studies are Appropriate for Investment Treaty Arbitration*

Scholars posit, "a basic premise of all empirical research—and indeed of every serious theory of inference—is that all conclusions are uncertain to a degree. After all, the facts we know are related to the facts we do not know but would like to know only by assumptions that we can never fully verify. The point is not to qualify every statement . . . but rather to estimate the degree of uncertainty inherent in each conclusion and to report this estimate along with every conclusion."[73] Prominent political scientists likewise explain that, while limited information and uncertainty is inevitable, "this uncertainty should not suggest that we avoid attempts at causal inference. Rather we should draw causal inferences where they seem appropriate but also provide the reader with the best and most honest estimate of the uncertainty of that inference. *It is appropriate to be bold in drawing causal inferences as long as we are cautious in detailing the uncertainty of the inference*." [74]

In international economic law, having only a small number of awards is not surprising. In international trade disputes, under the GATT era settlement processes, there were approximately nine cases a year; but under the new Dispute Settlement Understanding at the World Trade Organization, there are

---

[72] King, et al., *supra* note 55, at 8.

[73] Epstein & King, *supra* note 25, at 50.

[74] King, et al., *supra* note 55, at 76 (emphasis added); *see also* MILDRED L. PATTEN, UNDERSTANDING RESEARCH METHODS: AN OVERVIEW OF THE ESSENTIALS 113 (7th ed. 2009).

now approximately thirty to thirty-five per year.[75] Professor Colares empirically evaluates trade disputes and has, for example, analyzed a dataset of forty-one NAFTA disputes.[76] Other recent empirical legal studies outside of international economic law have used samples in the order of forty to sixty.[77] The *Harvard International* research is thus not unusual or improper.

B. Population Parameters and Limitations of Inferences

The research in the *Harvard International* article was done on a small population of awards at a certain moment in time, and drawing inferences on the population was technically unnecessary.[78]This means, based upon the population at issue (namely publicly available awards before June 1, 2006), for the models analyzed, the results were descriptively conclusive as to *that* population. The results were also the first efforts to hypothesize probabilistically about the population beyond initial dataset. This means, when using statistics to infer beyond the dataset (namely private awards or present and future arbitration outcomes), there is a risk of error. The *Harvard International* article therefore called for replication to assess whether the results are confirmed or rejected as ongoing population parameters. The article reported *p*-values, means, standard deviations and cell counts; it calculated effect sizes and conducted both *post hoc* and *a priori* power analyses (i.e. to identify the size of future samples required to ascertain even the smallest effect) to assess transparently the risk of error.[79] The objective was to identify what was known about the then-population and understand potential risks (and implicit limitations) related to the growing population.

As the *Harvard International* article disclosed, the statistical power of the inferences was low, and the related error rates were beyond levels traditionally

---

[75] Eric A. Posner & John C. Yoo, *Judicial Independence and International* Tribunals, 93 CALIF. L. REV. 1, 46 (2005); Robert E. Hudec, *The New WTO Dispute Settlement Procedure: An Overview of the First Three Years*, 8 MINN. J. GLOBAL TRADE 1, 15-16 (1999); *see also* Eric Reinhardt, *Adjudication without Enforcement in GATT Disputes*, 45 J. CONFLICT RESOL. 174, 176-77 (2001) (referring to a database created for the forty-eight year period between 1948-1994 that contained a total of 298 disputes resulting in 143 rulings, for an average of 6.48 disputes per year and 3.12 rulings per year).

[76] Juscelino F. Colares, *An Empirical Examination of Product and Litigant-Specific Theories for the Divergence Between NAFTA Chapter 19 and US Judicial Review*, 42 J. WORLD TRADE L. 691, 17 (2008).

[77] *See* George S. Geis, *An Empirical Examination of Business Outsourcing Transactions*, 96 VIRGINIA L. REV. 241, 257-58 (2010) (using a sample of sixty contracts); Jed H. Shugerman, *The Twist of Long Terms: Judicial Elections, Role Fidelity and American Tort Law*, 98 GEORGETOWN L.J. 1359, 1421 (2010) (using an initial sample of 42 and then other samples of 31, 14 and 11).

[78] BRUCE J. CHALMER, UNDERSTANDING STATISTICS 2 (1986).

[79] This comports with the norm that researchers "estimate the degree of uncertainty inherent in each conclusion and to report this estimate along with every conclusion". Epstein & King, *supra* note 25, at 52.

accepted levels.[80] Yet, while the low power is not ideal, it was reality. The reality was an inevitable by-product of what is, even now, arguably a small population that is evolving from a historically recent paradigm shift in international law. The *Yearbook Contribution* claims that, "Franck's statements of findings and conclusions were not qualified . . . by the lack of data".[81] The *Harvard International* article, however, neither hid the relatively low power of the analyses nor the risk of error; rather, the article stated these limitations clearly—and at regular intervals.[82] It acknowledged the existing archival data and its limitations.[83] The article included statements such as, "Out of extra caution in this sensitive area, it would therefore be prudent to engage in more research, with a larger sample, before establishing a population parameter that development status is not reliably associated with outcome."[84] For quantitative research in psychology, most journal articles include a paragraph describing the limitations of a study within the discussion section.[85] The *Harvard International* article went beyond a simple paragraph. It included a caveat at the end of each section of the research results as well as an entire section dedicated to the limitations of the article. Appendix I identifies over thirty statements in the *Harvard International* article alone that pertained to disclosures about the potential limitations.

On multiple occasions in the primary text,[86] the *Harvard International* article referenced the effect size (i.e., how statistically large the effect of development might be on outcome) and suggested that analyses were underpowered and required replication. Every model analyzed in the *Harvard*

---

[80] Cohen promulgated the standard of 80% power (20% risk of Type II error). Cohen, *supra* note 67, at 56.

[81] Van Harten, YB contribution at 17.

[82] In conformity with the *Rules of Inference* standard, *supra* note 73, that each research result should also reflect its limitation, every time the *Harvard International* article reported a result, it also identified the effect size, the power, the risk of error and the size of a future sample. Franck, *Harvard International* article, *supra* note 3, at 461, 461, 466-67, 469-70, 474-76. The *Yearbook Contribution* agrees that it was "rigorous" to calculate the error rates "and to indicate this limitation". Van Harten, YB contribution at 18.

[83] Franck, *Harvard International* article, *supra* note 3, at 458 n.105 (referring to Franck, *Evaluating Claims*, *supra* note 39, which in turn describes in detail the creation and nature of the archival data as well as related limitations, particularly pages 16-26).

[84] Franck, *Harvard International* article, *supra* note 3, at 472.

[85] JOHN J. SHAUGHNESSY, EUGENE B. ZECHMEISTER, & JEANNE S. ZECHMEISTER, RESEARCH METHODS IN PSYCHOLOGY 465 (7th ed. 2006).

[86] Although we are unaware of any guidelines for U.S. law reviews, the American Psychological Association (APA) has standards for reporting quantitative results. APA standards may not all be fully applicable to law—and it is questionable whether law reviews will understand, want or encourage scholars to include such information in above the line text. Yet the APA requires disclosure of information, such as statistical power, but does not require it to be placed in a certain location. APA Manual, *supra* note 47, at 30. The *Harvard International* article placed material in both the primary text and footnotes.

International article—each of which failed to identify any statistically significant reliable relationship between the variables of interest and outcome in the pre-2007 population—methodically considered the risk of error, which the *Yearbook Contribution* acknowledged.[87]

1. The *Harvard International* article did this for the OECD Chi-Square analysis of winners and losers, even though the sizes of the potential effects of development on outcome were either statistically less than small ($r = 0.04$) or small ($r = 0.14$).[88]

2. It did this for the World Bank Chi-Square analysis finding potential effects ranged from small to nearly medium ($r = 0.25$ and 0.24).[89]

3. It also did this for the OECD ANOVA, even though two of the effect sizes were less than small ($r = 0.01$) and the third effect (for the effect of a presiding arbitrator) was a bit more than small ($r = 0.14$).[90]

4. Finally, it did this for the World Bank ANOVA and found a less than small effect for the status of the presiding arbitrator ($r = 0.09$), but closer to medium potential effects for the respondent's development status ($r = 0.29$) and the interaction ($r = 0.19$).[91]

Although the results from every single model failed to reject the null hypothesis—namely that there was no link between outcome and development at the macro level—statistics offer lawyers a unique gift. Namely, even when we cannot find a statistically significant effect, presuming that the effect might actually exist, quantitative analysis permits the estimation of the *magnitude* of a potential effect. This is done with effect sizes, which describe the relative importance of an effect (i.e., how closely the variables are likely associated).[92] In other words, even if the analyses cannot demonstrate a reliable relationship among variables, it is still possible to estimate—presuming the effect exists—how big the effect is likely to be.[93]

---

[87] Van Harten, YB Contribution, at 18.

[88] Franck, *Harvard International* article, *supra* note 3, at 461.

[89] *Id.* at 463-64.

[90] *Id.* at 466-67. There were similar effect for raw data (interaction: $r = 0.17$; main effect of OECD status of respondent: $r = 0.01$; main effect for OECD status of presiding arbitration: $r = 0.01$); and there was also a disclosure about the possibility that the interaction was underpowered. *Id*. at 466 n.144.

[91] *Id.* at 469-70. There were similar effect sizes for the raw data; and these were all disclosed (interaction: $r = 0.22$; main effect of World Bank status of respondent: $r = 0.12$; main effect for World Bank status of presiding arbitration: $r = 0.04$). *Id*. at 470 nn.156-58.

[92] Lawless, et al., *supra* note 16, at 242.

[93] Brett Myors, *Statistical Power,* in The PSYCHOLOGY RESEARCH HANDBOOK: A GUIDE FOR GRADUATE STUDENTS AND RESEARCH ASSISTANTS 161, 163 (Frederick T. L. Leong &

None of the effect sizes in the *Harvard International* article suggested that development variables were likely to have a statistically large[94] effect on outcome. Moreover, only two models bordered on medium effects. The remainder of the analyses clustered around the possibility that development-related variables may have a minimal impact on outcome. The failure to appreciate the information from effect sizes and the isolated focus on significance tests means that the *Yearbook Contribution* ignored a key point. Even if a larger sample might detect a reliable link between outcome and development related variables, the impact is likely to be statistically small.[95]

C. Recognizing Results, Appreciating Limitations and Disagreeing with Inferences

Should one wish to disagree with the *Harvard International* article's interpretation of the results, one may do so. Likewise, if one disagrees with the normative suggestions that derive from those analyses—as suggested by a Public Statement that asserts investment treaty arbitration is not "fair, [and] independent" and advocates that governments should "refus[e] to pay arbitration awards against them"[96]—we welcome the debate since this is the value of academic freedom. No amount of normative evaluation or critique, however, can change the underlying findings, and no amount of wishing that the results were different can make them different. No matter how much one might dislike or be surprised by the results, the evidence should not be ignored. Disagreements about the results should be grounded in data and analysis that is as transparent, reliable, valid and replicable as the original.

In the case of the *Harvard International* article, the results did not support the hypothesis that there was a meaningful difference in outcome for the pre-2007 population based upon the variables studied. The probabilistic inference based upon those results was that there was likely also no effect in the current population. To flip this slightly, even if the operating research hypothesis was that differences in development status *should* be related to differences in outcome (a normatively troubling assumption), that research hypothesis was

---

James T. Austin eds., 2d ed., 2006); Franck, *Harvard International* article, *supra* note 3, at 457-58.

[94] This is Cohen's convention to demarcate effect sizes. *See* Cohen, *supra* note 67; *see also* ROPERT P. ABELSON, STATISTICS AS PRINCIPLED ARGUMENT 12, 46 (1995).

[95] There is a possibility that even "small" effects might make a difference—a second of time may affect whether an Olympic athlete receives a medal or not. Future research can consider how to justify effect size distinctions over time. For research in a new area, it is appropriate to work with established conventions until a different baseline is warranted.

[96] Gus Van Harten & David Schneiderman, *Public Statement on the International Investment Regime*, Aug. 31, 2010, ¶ 8, http://www.osgoode.yorku.ca/public_statement/documents/Public%20Statement.pdf.

not supported by the data. It may be cognitively easier to hold negative beliefs about investment arbitration, as cognitive biases[97] can increase the likelihood that people will recall recent and negative behavior—particularly negative and extreme experiences.[98] Yet cognitive preferences do not mean that assessments based on aggregate data are wrong.

Certainly there may be unfairness in specific cases, which is precisely why there are opportunities to challenge arbitrators for bias and why states may find it constructive to consider the net value of investment treaty arbitration on a country-specific basis. But, in our view, the case for rejection of the entire system was unwarranted on the basis of the evidence available in the *Harvard International* article. As the population grows and analyses evolve, this conclusion is a legitimate basis for on-going consideration.[99] If one wishes to set aside the current evidence because it does not comport with one's cognitive framework, that is a choice. We prefer to acknowledge the data we have, evaluate the results in light of their limitations, and continue to look for more data in the search for answers in an ever-shifting world.

IV. OPPORTUNITIES FOR THE FUTURE

In an effort to develop the academic discourse, this Section explores opportunities for a constructive discussion of social science methodologies. The authors are pleased that the *Yearbook Contribution* takes these issues seriously. Given issues we have raised, we highlight areas for future consideration related to data collection, measurement, researcher disclosures and potential insights from synthesizing quantitative and qualitative insights.

First, the *Yearbook Contribution* concurs with our assessment of the need for additional data. Nevertheless, systematic and careful data collection comes at a remarkable cost. It would be helpful to search for independent funding in support of research. The creation of the pre-2007 dataset was a by-product of a small seed grant by the University of Nebraska Law College and hundreds of hours of sweat equity invested by the researcher and research assistants. There

---

[97] *See supra* notes 9-10.

[98] Solomon E. Asch, *Forming Impressions of Personality*. 41 J. ABNORMAL & SOC. PSYCH. 258 (1946); *see also* Tiffany A. Ito et al., *Negative Information Weighs More Heavily on the Brain: The Negativity Bias in Evaluative Categorizations*. 75 J. OF PERSONALITY AND SOC PSYCH 887 (1998); Elizabeth A. Kensinger, et al., *Memory for Specific Visual Details can be Enhanced by Negative Arousing Content*. 54 J. OF MEMORY AND LANGUAGE 99 (2006); Lara G. Chepenick, et al*., The Influence of Sad Mood on Cognition*, 7 EMOTION 802, 807 (2007); Carrie L. Wyland & Joseph P. Forgas, *On Bad Mood and White Bears: The Effects of Mood State on Ability to Suppress Unwanted Thoughts*. 21 COGNITION & EMOTION 1513, 1518–19 (2007).

[99] There may be normative reasons for targeted reform related to determinacy, coherence, consistency and predictability of the substantive law. *See, e.g.,* Susan D. Franck, *The Legitimacy Crisis in Investment Treaty Arbitration: Privatizing Public International Law Through Inconsistent Decisions*, 73 FORD. L. REV. 1521 (2005).

is value in exploring how to pool resources in the creation of data to create a useful public good. In the interim, there is value in researchers sharing data, subject to special proprietary concerns, for legitimate research purposes.[100] Beyond a single repository for data, there may be value in exploring common protocols for the collection of data, even data from different sources.

Second, there is value in constructing and refining variables to improve the quality of models and analysis. For example, the *Harvard International* article articulated that it may be constructive to control for "the quality of expert evidence, the nature and scope of legal representation, and submissions by amicus curiae. Other variables affecting results may, however, be intrinsically tied to arbitration, such as the qualities and experiences of arbitrators."[101] Other refinements might be made, such as possible improvements we identified for the OECD variable. Likewise, there is value in a thoughtful discussion about creating a better proxy for the development status of a tribunal. The literature supported an evaluation of presiding arbitrators;[102] yet creating a composite development score for the tribunal (to reflect the overall variance) has challenges. Our hope is that thoughtful discussion about the costs and benefits of different variables will aid the future assessment of links between development and outcome.

Third, to enhance quality and to contextualize research, there may be value in disclosing more information.[103] In conformity with social science practices, there should be disclosure of operational definitions for measures, explanations of models, reporting of test results, cell counts, statements of standard deviations and effect sizes, *post hoc* power analyses and *a priori* power analyses. Other disclosures might relate to the identity of the researcher(s) and assistants and motivation for conducting research. Franck disclosed the identity of other data coders[104] and her own background.[105] In the interests of full disclosure, the dataset was created to test an implicit

---

[100] *See* APA Manual, *supra* note 47, at 12-13 (articulating issues for sharing data).

[101] Franck, *Harvard International* article, *supra* note 3, at 487-88. The two articles suggest an agreed need to explore the background of arbitrators in greater detail. *See* Van Harten, YB Contribution at 21 (referring to "a common culture or industry of arbitrators.").

[102] Franck, *Harvard International* article, *supra* note 3, at 448-53; *see also id.* at 457 ("This research is the first step in isolating variables linked with arbitrators' decisions").

[103] Disclosures of additional information about researchers may prove helpful in "guard[ing] against biases they may inadvertently introduce" to the research. Epstein & King, *supra* note 25, at 113-14.

[104] Franck, *Evaluating Claims*, *supra* note 39, at 1, 23. One of the co-authors of this article, Jenna Perkins, was the research assistant involved in initial data analysis that led to the more sophisticated models in the *Harvard International* article. *Considering Recalibration of International Investment Agreements: Empirical Insights*, *in* THE EVOLVING INTERNATIONAL INVESTMENT REGIME: EXPECTATIONS, REALITIES, OPTIONS (Eds. Jose E. Alvarez, et al. 2011).

[105] Franck, *Empiricism and International Law*, *supra* note 2, at 788 n.94. More information about Prof. Franck is available on her Washington & Lee faculty web page.

hypothesis in the *Legitimacy Crisis*, namely an instinct that tribunals can and would shift costs against parties who brought unmeritorious claims or defenses.[106] The answer, based upon the same dataset the *Yearbook Contribution* critiques, contradicted the research hypothesis; and this has been acknowledged.[107] Other research, using a different dataset from 2008-09, replicated key aspects of that costs scholarship.[108] The construction of the dataset was not designed to prove—or disprove—the presence of bias (or even assess perceived bias). It was to assess fiscal costs in arbitration and justifications for cost outcomes. The dataset, once created, provided the opportunity to address other issues of public concern.

Finally, a constructive by-product of the current debate would be a dialogue about the net costs and benefits of investment treaties and dispute resolution mechanisms. In other words, blending insights from research models that use quantitative and qualitative methods may prove fruitful. One reasonable suggestion is that, not only are such insights beneficial, but they are necessary for the considered evolution of policy.

The *Yearbook Contribution* seems torn about the current value of quantitative analyses of investment arbitration.[109] As generating sufficient power to identify reliably even the small effects could require a wait of potentially 50 years, we believe it unwise to exclude potential insights. Quantitative scholarship offers a lens to help us understand reality more systematically, hopefully free from cognitive biases that may disrupt our capacity to rationally process information. But quantitative data in isolation may create inadvertent blind spots that necessitate recognition of qualitative experiences. Holistic understanding of a system's operation using aggregate data can provide insights for systemic assessment. Recognizing the system continues to grow, the existing data and analysis did not support the conclusion that there was a fundamental flaw in investment treaty arbitration, at least as a function of the development status variables. But likewise, the analyses from aggregate data did not necessarily mean that the investment arbitration system is perfect or without flaw. Although based upon the data, variables and models, the outcomes did not indicate systemic differences in outcomes, there may be individual cases that are of concern. For example,

---

[106] Franck, *Legitimacy Crisis*, *supra* note 99, at 1591-92, 1603, 1622.

[107] Susan D. Franck, *Rationalizing Costs in Investment Treaty Arbitration*, 88 WASH. U. L. REV. 769 (2011).

[108] David Smith, Note, *Shifting Sands: Cost-and-Fee Allocation in International Investment Arbitration*, 51 VA. J. INT'L L. 749, 752-53, 756 (2010).

[109] The *Yearbook Contribution* states "lack of data on awards in investment arbitration makes it especially hazardous, and often impossible, to infer conclusions about bias based on quantitative research" and later continues that "there are at present major limitations to the use of quantitative methods to test for bias in investment arbitration". Van Harten, YB Contribution at 2, 34. Yet it also caveats that this is "not a criticism of the use of quantitative methods *per se*" and suggests "a range of qualitative and quantitative methods may be used to examine expectations of bias". *Id.* at 23, 1.

Franck previously critiqued an award against Ecuador given the scope of the tribunal's substantive legal determination.[110] One might imagine the specific experiences of a repeat-player respondent may be an appropriate basis for an individualized determination—in light of the broader framework—about whether the risks and costs of dispute resolution are worth the benefits brought from a treaty. Whether the system as a whole reflects those same concerns, however, is another matter.

V. CONCLUSION

Concerns about methodology are an integral element of academic discussion. Scholars can, and should, take methodological concerns seriously. We can all benefit from constructive discussions that recognize social science norms and work to advance the evolution of knowledge through dialogue. In order to make us better producers of and consumers of research, that conversation should occur with an understanding of social science norms. For this reason, it is helpful to situate the *Harvard International* article and the *Yearbook Contribution* within those norms.

First, the *Yearbook Contribution* suggests that use of OECD membership to analyze development status created unacknowledged limitations. The *Harvard International* article expressly defined the measure, identified those countries that were OECD members, recognized limitations of the OECD measurement and cross-checked the variable using a four-category World Bank variable. Even using Van Harten's definition, the research results did not change. All the results were the same. At the macro level, it was not possible to prove a statistical link between development and outcomes of investment treaty arbitration.

Second, the *Yearbook Contribution* suggests there was insufficient data (i.e. awards) to conduct quantitative research. The research was conducted on an entire population of public awards available before June 1, 2006; and it was and continues to be proper to make statistical conclusions about that population. Whether the results are replicable over time and applicable to both public and private awards, is another matter. As research evolves, additional evidence might increase our confidence in the results from the *Harvard International* article. It is also possible that future research may reject or suggest additional subtleties about the baseline findings in the *Harvard International* article. In the interim, with is a growing population and state sovereignty is at stake, systematically gathered and compiled data—as limited as it is—is normatively preferable to relying upon other forms of knowledge

[110] Susan D. Franck, *International Decisions:* Occidental Exploration and Production Company v. The Republic of Ecuador, 99 AM. J. INT'L L. 675 (2005).

including intuition, unverified statements by authorities or chance alone.

Third, the *Harvard International* article offered transparent research methodology and identified limitations of the research. It articulated the measures, models, research results, sample sizes, standard deviations, effect sizes, *post hoc* power analyses and *a priori* power analyses. In more than thirty places—including a three-page section entitled "Understanding the Limitations of the Analysis"—the article discussed its limitations.[111] The proprietary primary dataset—with approximately 200 different variables and 20,000 pieces of data—has always been available upon request from the *North Carolina Law Review*. Upon request, data was provided to other researchers, including Professor Van Harten. In light of the reported results, transparency of methodology and acknowledgement of limitations, the *Yearbook Contribution*'s claims of "overreaching" is surprising. As every model failed to find a link between development and outcome, the *Harvard International* article's results were framed in terms of "temper[ing] this cautious optimism properly" to recognize the limitations and need for replication.[112]

Finally, the *Yearbook Contribution* appears concerned the *Harvard International* article did not address the appearance of bias. It is correct that the research did not address perceived bias. As initial research in a new area, the *Harvard International* article instead focused on actual bias by evaluating whether there were systemic or unexpected differences in outcomes across categories. Although variables about perception are often subjective, by using valid and reliable measures, future research might constructively explore indicators of perceived bias. For example, in an effort to explore both perceived and actual bias, it may be fruitful to explore data on arbitrator challenges, which are derived from arbitration rules designed to address problems of bias and partiality in actual disputes.

In an effort to take a serious subject seriously, we have sought to identify opportunities for the future and to generate a dialogue about improving data collection, variables for analysis and disclosure of information. Our hope is, having recognized that careful research and informed critiques are prudent in international economic activity, that we can move forward constructively—with a better understanding of social science norms—in an effort to engage in quality research and an informed normative debate.

---

[111] *See* Appendix I.

[112] Franck, *Harvard International* article, *supra* note 3, at 478; *see also id.* at 473, 487.

RESPONSE TO REPLY

Susan D. Franck, Calvin P. Garbin and Jenna M. Perkins

In *Through the Looking Glass*, we explored fundamental social science norms and considered their application to investment treaty arbitration in order to evaluate the concerns in Professor Van Harten's *Yearbook Contribution*[113] related to a *Harvard International Law Journal* article.[114] In light of the *Reply* and given our belief that dialogue is a productive way to advance knowledge, we now highlight the areas of agreement, clarify our understanding of "bias" and consider the limitations of inferences from statistical results. We conclude by emphasizing that all research methods—whether empirical or otherwise—have limits, but the inevitable limitations that come with a quantitative evaluation of international investment law should not be invoked to inhibit serious, careful and methodologically sound research.

There is important common ground among all of the authors involved in the dialogue. First, all authors appear to agree that research of investment treaty arbitration is both relevant[115] and worthy of ongoing inquiry.[116] Second, all authors agree that construct validity—how we define social constructs, including the meaning of Development Status—is worthy of thoughtful discussion.[117] We cannot, will not and did not suggest that the definition of OECD is the exclusive method for evaluating Development Status. This was why the *Harvard International* article identified concerns related to Mexico and replicated the analysis using a different measure, namely the World Bank's own classification.[118] The bottom line is, irrespective of how we defined a respondent's Development Status, at the macro level there was no statistically significant difference between the outcomes for different types of

---

[113] Gus Van Harten, *The Use of Quantitative Methods to Examine Possible Bias in Investment Arbitration*, *supra* YEARBOOK OF INTERNATIONAL INVESTMENT LAW [hereinafter *Yearbook Contribution*].

[114] Susan D. Franck, *Development and Outcomes of Investment Treaty Arbitration*, 50 HARV. INT'L L.J. 435 (2009) [hereinafter *HILJ Article*].

[115] *Yearbook Contribution*, *supra* note 113, at 3; Susan D. Franck, Calvin P. Garbin & Jenna M. Perkins, *Through the Looking Glass: Understanding Social Science Norms for Analyzing International Investment Law*, *supra* YEARBOOK OF INTERNATIONAL INVESTMENT LAW, at 5 n.26 [hereinafter *Looking Glass*].

[116] *Yearbook Contribution*, *supra* note 113, at 10 (stating that "[a]n advantage of empirical study is its potential to uncover facts that allow existing theories to be refined or revised"); *Looking Glass, supra* note 115, at 5.

[117] *Yearbook Contribution*, *supra* note 113, at 13-14; *Looking Glass*, *supra* note 115, at 7.

[118] *HILJ Article*, *supra* note 3, at 455 n.109 (discussing issues related to Mexico); *id.* 462-64, 467-70 (analyzing Development Status using the World Bank's classification).

respondent states. Third, all authors appear to agree that there is no reliable empirical evidence, either in the *Harvard International* article or elsewhere, that the international investment law system exhibits actual bias in arbitration outcomes against developing countries. Professor Van Harten has not provided any holistic empirical analysis, with sound *ex ante* research methodology, demonstrating that investment treaty arbitration outcomes exhibit a systemic bias against developing countries. Evidence of the presence or absence of actual bias is, of course, important in evaluating whether perceptions of bias are reasonable in the circumstances.

To that end, it may be constructive to clarify our definition of "bias". For the purposes of quantitative analysis, bias involves considering whether—during the critical process of rendering an arbitral award—outcomes exhibited systemic differences for variables of interest. The objective was to look at actual awards to search for verifiable indicators of bias in real case outcomes. The *Harvard International* article and *Through the Looking Glass* used significance testing (and the intertwined consideration of effect sizes) to explore whether there were meaningful and reliable differences in outcomes across different development-based categories. Put in a more colloquial way, by narrowing the empirical microscope to focus on development, the research explored whether Development Status affected outcomes such that investment treaty arbitration was the statistical equivalent of tossing a two-headed coin.

Hoping that rule of law adjudicative systems do not create rigged outcomes, the research hypothesis was the same as the null hypothesis, namely: there would be no reliable relationship between development variables and outcome. The research hypothesis also could have taken one of two different forms: (1) there would be a systemic difference in arbitration outcomes that worked to the *detriment of the developed world*, or (2) there would be a systemic difference in arbitration outcomes that worked to the *detriment of the developing world*. Even if one of those two alternatives had been adopted as the research hypothesis, the results of the statistical tests would have all been the same. For every single one of the twenty-four models analyzed in the *Harvard International* article and *Through the Looking Glass*, there was no evidence of substantial or significant statistical differences at the macro level for Development Status. Even if we had asserted that the system exhibited bias, the results would not have supported that position.

As with other forms of knowledge acquisition, making inferences from statistical evidence inevitably involves risk. Professor Van Harten correctly says that the failure to find lack of bias in one model does not necessarily mean that investment treaty arbitration is completely free of bias. Inferences can only be made about the variables analyzed, like the development variables we explored. It is also notoriously difficult to prove a negative. Statistical tests, designed to provide reliable evidence of difference, cannot conclusively prove equivalence; but they offer useful evidence about the possibility of a lack of a difference.

This makes Van Harten's analogy apt, namely: looking for bias in investment treaty arbitration may be the equivalent of looking for pliers in a messy garage.[119] Both situations are complex and may cause frustration. One may wish to find the object one desires, whether a pair of pliers in a garage or bias in investment treaty arbitration. But when one explores the traditional locations where a reasonable observer would expect to find the pliers—such as the tool bench where the pliers were last used, in the drawer where the pliers are usually stored, or on the floor in between the bench and the drawer—that may be useful initial evidence that the pliers are likely not in the garage. Likewise, when research explores the most obvious aspects of bias (such as bias against the developing world claimed by a head of state) and it cannot be found, that is valuable evidence. While we cannot definitively claim to know about the location of pliers (or the existence of bias), the empirical enquiry can begin to minimize the uncertainty.

All models of knowledge are inevitably limited, and empirical methodologies are no exception. While we can never eliminate all risk of error given the probabilistic nature of empirical enquiry, we can increase our confidence in the results through replication. The research initiated thus far is a starting point—not an ending point. Social scientists and methodologists spend major portions of their professional careers debating research, seeking ways to improve measures and models and considering the finer points of inference. It is a pleasure to see that empirical research on investment arbitration has already reached this advanced stage even at such a nascent phase of its evolution. This discussion has been an illuminating opportunity to consider ways to improve upon initial good-faith research efforts designed to answer an important question about international investment law. Professor Van Harten has usefully pointed out the complexity of measuring constructs and of processing the data to produce a meaningful and trustworthy conclusion, especially given the limited data available. However, it is important to note the convergence of the results from the many alternatives that were explored. At present, from these analyses of these data, there is no consistent evidence supporting a conclusion of substantial bias in arbitration decisions. Empirically based discussions of consistent and substantial research should continue, both with additional analysis of the pre-2007 data and additional data from the evolving population. The issues at stake are too important to do anything less.

---

[119] Gus Van Harten, *Reply*, *supra* YEARBOOK OF INTERNATIONAL INVESTMENT LAW, at 14, *citing* David F. Parkhurst, *Statistical Significance Tests: Equivalence and Reverse Tests Should Reduce Misinterpretation*, 51 BIOSCIENCE 1051, 1053 (2001).

Appendix I

References to Research Limitations in Franck (2009)

1. "Part V describes the initial results and explains the research limitations."[120]

2. "It cautions, however, that proper contextualization and replication of the research is necessary."[121]

3. "Part V.E then describes the limitations of the research."[122]

4. "Third, recognizing the limited nature of the inferences, the research can provide information to stakeholders, such as government officials negotiating treaties, who may wish to consider the potential implications for the design of their dispute resolution systems."[123]

5. "Recognizing the limitations of the research, this Part argues that reform has the benefit of promoting procedural legitimacy by addressing concerns related to perceptions about the system's fairness."[124]

6. "By measuring the potential strength of a relationship between two variables, effect sizes aid assessment of whether, on a normative level, the size of a reliable statistical difference is a matter of practical concern or is so tiny as to be irrelevant."[125]

7. "Finally, section E describes the limitations of the analyses and related inferences."[126]

8. "The underlying data has limitations given that it comes from publicly available archives and dates only to June 1, 2006. There are now approximately three more years of data to collect and analyze. It is necessary to acknowledge this limitation, and future research should replicate the analyses."[127]

---

[120] Susan D. Franck, *Development and Outcomes of Investment Treaty Arbitration*, 50 HARV. INT'L L.J. 435, 439 (2009).
[121] *Id.* at 440.
[122] *Id.*
[123] *Id.*
[124] *Id.*
[125] *Id.* at 458.
[126] *Id.*
[127] *Id.* at 459 n.129.

9. "As the two statistically significant pairwise comparisons involve these countries, inferences from the data are limited."[128]

10. "The effect sizes also suggest that further research is warranted. Many analyses did not achieve statistical significance and had effect sizes that were so tiny as to be of little practical effect."[129]

11. "Out of extra caution in this sensitive area, it would therefore be prudent to engage in more research, with a larger sample, before establishing a population parameter that development status is not reliably associated with outcome."[130]

12. "Given that the data did not suggest a reliable link between development status and outcome, the evidence begins to suggest that the investment treaty arbitration system appears to be functioning reasonably well at the macro level."[131]

13. "The presence of two statistically significant simple effects in the amounts awarded, however, suggests tempering this 'good news,' as certain permutations merit further reflection."[132]

14. "First and foremost, further research is necessary to replicate these findings to assess whether they are real population parameters or the result of statistical chance."[133]

15. "First, there may be limitations to the strength of the inferences, as they may not reflect population parameters."[134]

16. "Replication with expanded data is necessary to avoid establishing a population parameter that may be due to chance alone."[135]

17. "Second, there may be issues about the validity of the statistical conclusions. Effect sizes suggest that the power of the research is

---

[128] *Id.* at 472.
[129] *Id.*
[130] *Id.*
[131] *Id.* at 473.
[132] *Id.*
[133] *Id.* at 474.
[134] *Id.*
[135] *Id.* at 475.

relatively low. It would be prudent to establish a broader pool of data, based on a priori power analysis, to confirm, clarify, contradict, or supplement these findings."[136]

18. "This means that there would be utility in replicating and expanding this research by using more complicated models and additional variables to refine both the research questions and the statistical conclusions."[137]

19. "Since there is now nearly three additional years' worth of data to gather and analyze, future research should replicate the analysis."[138]

20. "Additional control variables could minimize the risk of statistical confounds. For example, given the limited and missing data in the present database, it was not possible to control for variables such as differences between amounts claimed versus amounts awarded. Future research might have enough data to usefully add this factor and consider other variables such as the number of arbitrators, the gender of arbitrators, the institutions administering the arbitrations, and the identity of lawyers representing the parties."[139]

21. "This makes future research challenging but does not diminish the importance of replication and convergence. As the data pool expands, analysis will be possible, but it is important to recognize the potential limitations of its statistical power."[140]

22. "The theme from the present research suggests creating targeted adjustments when there is a valid and reliable diagnostic demonstrating the value of such modifications, while recognizing the possible limitations when implementing policy changes."[141]

23. "Despite this general cautiously good news, however, there are areas for improvement and prudence. Section B therefore explores various opportunities stakeholders may take to improve investment treaty dispute resolution, and section C suggests areas of caution so that counsel and arbitrators can take even greater care during the adjudicative process to promote procedural justice."[142]

---

[136] *Id.*
[137] *Id.*
[138] *Id.* at 475.
[139] *Id.* at 475 n.176.
[140] *Id.* at 476.
[141] *Id.* at 476-77.
[142] *Id.* at 477.

24. "The results suggest that it is necessary to temper this cautious optimism properly. Empirical diagnosis found two simple effects where presiding arbitrators from the developing world made larger awards against developing countries and smaller awards against developed countries. Also, while not statistically significant and potentially affected by statistical outliers, there was a phenomenon whereby tribunals with arbitrators from developed countries rendered higher awards against respondents from the developed world, whereas tribunals with presiding arbitrators from the developing world rendered higher awards against countries from the developing world."[143]

25. "While they bear watching and are worthy of replication, the initial results suggest that there may be areas that require—or would simply benefit from—targeted interventions to improve the design of future dispute resolution systems."[144]

26. "First, it means that there are reasonable concerns about the reliability of the statistical conclusions—namely that tribunals with presiding arbitrators from Middle Income countries treat countries differently on the basis of their development background."[145]

27. "This research, particularly if replicated, creates a strong case for the creation of an appellate body or even a stand-alone investment court."[146]

28. "The notion that outcome is not associated with arbitrator or respondent development status should be a basis for cautious optimism. It provides evidence about the integrity of arbitration and casts doubt on the assumption that arbitrators from developed states show a bias in terms of arbitration outcomes or that the development status of respondent states affect such outcomes." [147]

29. "The lack of a reliable relationship between development status and out- come suggests that other variables or combinations of variables may drive arbitration results. Some of these variables may be completely disassociated from the arbitration process. Possible

---

[143] *Id.* at 478.
[144] *Id.*
[145] *Id.* at 479.
[146] *Id,* at 484.
[147] *Id,* at 487.

> variables could include those traditionally associated with neutral, adjudicative forums, whether courts, claims, commissions, or arbitrations, such as the quality of expert evidence, the nature and scope of legal representation, and submissions by amicus curiae. . . . Future research might usefully assess the impact of these and other variables in order to gain a more nuanced understanding of factors that are reliably associated with outcome. "[148]

30. "While the general initial results are encouraging, one should contextualize them properly, given their limitations."[149]

31. "Even if the results are not replicable . . . the critical message from the initial results is clear, namely that more empirical research is needed to examine development issues in greater detail and consider how best to enhance the integrity of the dispute resolution process."[150]

32. "This Article suggests that while there may be some problems with arbitration, it is not clear that a development divide affects outcomes. However, this is subject to evolution based upon new research." [151]

---

[148] *Id,* at 487-88.
[149] *Id.* at 488.
[150] *Id.*
[151] *Id.*

Appendix II: Adjusting for Transformations and Net Present Value in Amounts Awarded

We are grateful to an anonymous reviewer who encouraged us to explore certain methodological concerns about ways to minimize skewing of data and the implications of the time value of money. This Appendix explores those technical considerations in detail.

I. TRANSFORMING DATA AND MINIMIZING SKEWNESS

The *Harvard International* article identified the need to run multiple versions of a model to ensure that the results were robust. It analyzed both raw and winzorized data.[152]

Winsorising is a standard practice in psychology research.[153] The objective of winsorizing is to increase the symmetry of a distribution, decrease the skewness and while retaining information about real cases within a dataset. It first requires isolating the 25th percentile or above the 75th percentile of the values in the dataset (commonly called Tukey's Hinges). Extreme cases are then identified using a calculation to identify the upper and lower bounds of the inter-quartile distribution.[154] Rather than eliminating the outliers, winsorizing reassigns the extreme values to a value that is the same as the closest non-extreme observation.[155] In the *Harvard International* article, this meant that for the seven awards that were deemed outliers, the "winsorized" value of the award was the equivalent of the dataset's upper-bound, namely US$5,675,537.50. All other awards retained their original value.

Prior to Franck conducting the analyses that lead to the *Harvard International* article, all of the authors agreed that using winsorized data

---

[152] *See* Susan D. Franck, *Development and Outcomes of Investment Treaty Arbitration*, 50 HARV. INT'L L.J. 435, 456 nn.116-17 (2009) (explaining the process of and choice to winsorize).

[153] *See e.g.,* David D. DiLalla & Stephen J. Dollinger, *Cleaning Up Data and Running Preliminary Analyses* 227, 248, *in* The PSYCHOLOGY RESEARCH HANDBOOK: A GUIDE FOR GRADUATE STUDENTS AND RESEARCH ASSISTANTS (Frederick T. L. Leong & James T. Austin eds., 2d ed., 2006); John W. Tukey, *The Future of Data Analysis*, 33 ANNALS MATH. STAT. 1, 18-19 (1962); *see also* RAND R. WILCOX, INTRODUCTION TO ROBUST ESTIMATION AND HYPOTHESIS TESTING 27-28 (2005).

[154] DAVID C. HOAGLIN, FREDERICK MOSTELLER & JOHN WILDER TUKEY, UNDERSTANDING ROBUST AND EXPLORATORY DATA ANALYSIS 38-41 (1983).

[155] *See e.g.,* W. J. Dixon, *Simplified Estimation from Censored Normal Samples*, 31 ANNALS MATH. STAT. 385 (1960); DAVID SHESKIN, HANDBOOK OF PARAMETRIC AND NONPARAMETRIC STATISTICAL PROCEDURES 404 (3d ed. 2004).

would be preferable to analyzing raw or trimmed data in isolation. For the population of investment arbitration awards, winsorizing data kept the amount awarded in a scale that people can understand (namely US dollar values), reduced skewing and retained real case data from the pre-2007 population. The skewness of amounts awarded in the pre-2007 dataset was: (1) raw data = 5.311; (2) trimmed data = 2.437; and (3) winsorized data = 1.414.[156] Although the raw data reflected real dollars awarded, its skewing (over 5.0) meant it could not be analyzed in isolation. The high skewing of the trimmed data also decreased its analytical value. The lower skewing (i.e. closer to 0.0) for the winsorized data meant that its use was preferable.

There are additional ways to minimize asymmetries including log, square root and inverse transformations.[157] The objective of the transformations is to obtain a skewness that is acceptably close to zero.[158] Although the transformed data is not in an easily understandable scale (i.e. not U.S. Dollars), the transformations exhibited various levels of skewness: (1) square root = 3.734, (2) log = 0.506, and (3) inverse = -0.404. As both the log and inverse transformations had values closer to zero, those transformations provide an opportunity to re-analyze the data and evaluate the robustness of the *Harvard International* models.

We replicated the analyses on page 465 of the *Harvard International* article to explore the OECD-based model using both the log and transformed data. Irrespective of whether log or inverse transformations were used, none of the analyses revealed any statistically significant effects. For example, using the log transformations, there was no statistically significant interaction among the OECD status of the respondent, the OECD status of the presiding arbitrator and the amounts tribunals awarded (F (1,45) = 0.070; p = 0.79; r = 0.04; n = 49).[159] Likewise, the log data did not identify a statistically

[156] The standard error for the skewness statistics was 0.330.

[157] Log transformation uses an algorithmic log conversion of the data; square root transformations use a square root transformation of the data; and inverse transformations involve using a function. *See, e.g.*, ROBERT M. LAWLESS, JENNIFER K. ROBBENNOLT & THOMAS S. ULEN, EMPIRICAL METHODS IN LAW 217-22 (2010); Sheshkin, *supra* note 155, at 403-06. Trimming is another method that identifies outliers but addresses them by removing them from the dataset (i.e. ignoring the existence of actual data). Lawless, *supra*, at 222.

[158] Lawless, *supra* note 157, at 221.

[159] For the inverse transformation, there was no significant interaction (F (1,45) = 0.329; p = 0.57; r = 0.09; n = 49). Given the less than small effect sizes for the interactions based upon the log (r = 0.03) and the inverse (r = 0.09), the power of those models was less than 0.20, which suggests a greater than 80% risk of error. If there is no statistically significant result in a hypothesis test and the effect size is < .10, the analysis may not suffer from a power problem (i.e. the sample is of an insufficient size to detect a significant effect) or the possible effect may be so small that it will be difficult to justify resources to research the issue. This means that to obtain the standard 0.80 power for effects of r = 0.10 or lower, (a) the OECD-based model requires a sample of 781, and (b) the World Bank-based model requires a sample of

significant relationship between the OECD status of the respondent and amounts awarded (F (1,45) = 0.030; p = 0.86; r = 0.03; n = 49).[160] Also, there was statistically significant relationship between having a tribunal with a presiding arbitrator from an OECD member and amounts awarded (F (1,45) = 1.280; p = 0.26; r = 0.17; n = 49).[161] None of the tests, which used data with minimal skewing, identified a statistically significant relationship between OECD-based variables and amounts awarded. Moreover, the possible size of the effects studied were either less than small or small.[162] The congruity in results—particularly in log transformed data and inversely transformed data with minimal skewing—demonstrates the OECD-based models and results in the *Harvard International* article were robust.

Likewise, we replicated the analyses on page 468 of the *Harvard International* article to consider whether the World Bank-based model exhibited different results. Once again, irrespective of whether log or inverse transformations were used, there were no statistically significant relationships. Using the log data, there was no statistically significant interaction with the World Bank status of the respondent, the World Bank classification of the presiding arbitrator and amounts awarded (F (3,41) = 0.72; p = 0.55; r = 0.22; n = 49).[163] The log data also failed to reveal a reliable relationship between the World Bank status of a respondent state and the amounts awarded (F (3,41) = 1.115; p = 0.35; r = 0.28; n = 49).[164] Similarly, the log data failed to

---

1,562 awards, both of which could take decades.

[160] For the inverse transformation, there was no main effect for respondent's OECD status (F (1,45) = 0.102; p = 0.75; r = 0.05; n = 49). Given the less than small effect sizes for the interactions based upon the log (r = 0.03) and the inverse (r = 0.05), the power of those models are less than 0.20, which suggests an 80% risk of error.

[161] For the inverse transformation, there was no significant main effect for the presiding arbitrator's OECD status (F (1,45) = 1.468; p = 0.23; r = 0.18; n = 49). The power of the model using log data was 0.20, and the power for the transformed data was 0.30. Given the less than small effect sizes for the interactions using the log (r = 0.03) and inverse (r = 0.05), the power of those models was less than 0.20, which suggests an 80% risk of error. Yet the results may not be a function of pure power given the less than small effect size. *See supra* note 159.

[162] Appendix II adopts the same definitions of small, medium and large effects that were used by Jacob Cohen and described in *Through the Looking Glass*, *supra* at 13, n.67.

[163] For the inverse transformation, there was no significant interaction (F (3,41) = 0.751; p = 0.53; r = 0.23; n = 49). Given the small to medium effect sizes for the interactions based upon the log (r = 0.22) and the inverse (r = 0.23), the power of those models was 0.30, which suggests a 70% risk of a Type II error and a need for replication before drawing definitive conclusions about the current population.

[164] For the inverse transformation, there was no significant main effect for the World Bank classification of the respondent (F (3,41) = 0.990; p = 0.41; r = 0.26; n = 49). The power of these analyses ranged from 0.40-0.50, which indicates a 50-60% risk of a Type II error. As the effect sizes bordered on medium, replication is warranted before drawing

show a reliable link between the World Bank classification of the presiding arbitrator and the amounts awarded (F (1,41) = 0.080; p = 0.78; r = 0.14; n = 49).[165] None of the World Bank based models were able to identify a statistically significant relationship between World Bank-based classifications and amounts awarded. Like the analyses on winsorized data in the *Harvard International* article, many of the effect sizes were small or less than small, but the effect for a respondent's World Bank status bordered on medium-sized. The congruity in results demonstrates the models and results in the *Harvard International* article were robust.

II. ADJUSTING FOR THE TIME VALUE OF MONEY

The *Harvard International* article analyzed the actual amounts awarded in a case using the raw amount awarded on the date of the award. Given the variations in the size of amounts awarded at different times, the *Harvard International* article did not adjust for inflation.

A reasonable research hypothesis is that even adjusting for time value of money, the results of the statistical analyses would not result in a meaningfully different outcome as, for those awards where the tribunals made an award in favor of the investor, there was a mix of awards over the dataset's time period. For example, in the five lowest awards when the tribunals made awards in favor of investors, there was a range of awards over time: (1) *Iurii Bogdanov, Agurdino-Invest Ltd. and Agurdino-Chimia JSC v. Republic of Moldova*: September 25, 2005 for US$24,603; (2) *Maffezini v. Kingdom of Spain*: November 13, 2000 for US$155,314; (3) *Pope & Talbot v. Canada*: May 31, 2002 for US$407,646; (4) *Asian Agricultural Products v. Sri Lanka* for US$460,000; and (5) *Fexdax v. Venezuela*: March 9, 1998 for US$598,950. Similarly, for the five highest awards in the dataset, there was a range of awards over time: (1) *American Manufacturing and Trading v. Zaire*: Feb 21, 1997 for US$9,000,000; (2) *Metalclad v. Mexico*: August 30, 2000 for $16,685,000; (3) *Occidental Exploration and Production v. Ecuador*, July 1, 2004 for US$71,533,549; (4) *CMS Gas Transmission v. Argentina*, May 12, 2005 for US$133,200,000; and (5) *CME Czech Republic BV v. Czech Republic*, March 14, 2003 for US$269,814,000.

In an effort to explore the empirical question of whether adjusting for

---

definitive conclusions about the lack of an effect in the current population.

[165] For the inverse transformation, there was no significant main effect on amounts awarded based upon the presiding arbitrator's World Bank classification (F (1,41) = 0.060; p = 0.81; r = 0.04; n = 49). Given the small effect size for the log (r = 0.14) and the less than small effect size for the inverse (r = 0.04) data, the power of the analyses was 0.20 or lower. For the small effect, this indicates an 80% risk of a Type II error and the need for replication before drawing conclusive inferences beyond the dataset studied.

inflation makes a difference, we first considered how best to adjust the actual amounts awarded. We decided to adjust the amounts awarded according to two different pre-existing indexes to see if derivative adjustment changed the results of the analyses. First, we selected the Consumer Price Index (CPI), which is a standard adjustment for inflation and based upon the U.S. Bureau of Labor Statistics. As investment arbitration is not necessarily about U.S. consumer prices, the second index we selected was based upon Gross Domestic Product (GDP) per capita. Using these two different indexes, we converted the amount of the award from the date it was made (awards ranged from the 1990 to 2006) into 2006 U.S. dollars.

Using these adjusted values, we replicated the analyses on pages 465 and 468 of the *Harvard International* article. None of the analyses—irrespective of whether CPI-adjusted or GDP-adjusted data was used—revealed a statistically significant difference in mean amounts awarded as a function of development-related variables studied. It did not matter whether an OECD or World Bank based model was used. It did not matter which dependent variable was used to consider the effects of skewing within the data. For every analysis we ran to adjust for inflation, there was no relationship between development variables and amounts awarded. Even with the limitation of low power, this demonstrates that the models and results in the *Harvard International* article were robust.

A. OECD Based Models: CPI and GDP Adjusted Amounts Awarded

We first considered the OECD-based models using the CPI-adjusted values for awards. Given the concerns identified above, we first considered the skewness of the CPI-adjusted data. To replicate the *Harvard International* article and include variables with decreased skewness, we selected: (1) raw CPI-adjusted, (2) winsorized CPI-adjusted, (3) log CPI-adjusted, and (4) inverse CPI-adjusted data for analysis.[166] None of these different models revealed a statistically significant link between the OECD-based constructs of development and the amounts awarded. First, using the log transformation for CPI-adjusted amounts, there was no interaction among the OECD status of the respondent state, the OECD status of the presiding arbitrator and amounts awarded (F (1,45) = 0.068; p = 0.80; r = 0.04; n = 49).[167] Second, there was

[166] The skewness of the CPI-adjusted amounts awarded was as follows: (1) raw = 5.421, (2) trimmed = 2.437, (3) winsorized = 1.424; (4) square root = 3.725, (5) log = 0.503, and (6) inverse = -0.404.

[167] The power of the interaction was less than 0.20. Given the non-significant results and less than small effect size, the failure to identify a significant result may not purely be a

no statistically significant relationship between being an OECD respondent and amounts awarded (F (1,45) = 0.028; p = 0.87; r = 0.02; n = 49).[168] Third, there was no statistically significant relationship between having a presiding arbitrator from an OECD country and the amounts awarded (F (1,45) = 1.296; p = 0.26; r = 0.17; n = 49).[169] The results from the raw CPI-adjusted data were similar.[170] Likewise, the results from the winsorized[171] CPI-adjusted data and the inverse transformation[172] of the CPI-adjusted data failed to reveal any statistically significant results.

---

function of power. *See supra* note 159.

[168] The power of the main effect of the respondent's OECD status was less than 0.20. Given the non-significant results and the less than small effect size, the failure to identify a statistically significant difference may not purely be a function of power. *See supra* note 159.

[169] The power of the main effect of the presiding arbitrator's OECD status was 0.20. Given 80% risk of a Type II error, replication is warranted before making strong inferences beyond the pre-2007 population.

[170] There was no statistically significant interaction with OECD status of the respondent, OECD status of the presiding arbitrator and amounts awarded (F (1,45) = 1.338; p = 0.25; r = 0.17; n = 49). There was no statistically significant relationship with the OECD status of the respondent and amounts awarded (F (1,45) < 0.001; p = 0.99; r = < 0.01; n = 49). There was also no statistically significant relationship between having a presiding arbitrator from an OECD state and amounts awarded (F (1,45) = 0.003; p = 0.96; r = 0.01; n = 49). For the interaction, with a small effect, low power of 0.20 and an 80% risk of a Type II error, replication is warranted before making conclusions regarding the current population. For the two main effects with less than small effect sizes (power less than 0.20), the failure to identify a statistically significant difference may not purely be a function of power. *See supra* note 159.

[171] For the winsorized data, there was no statistically significant interaction with OECD status of the respondent, OECD status of the presiding arbitrator and amounts awarded (F (1,45) = 0.003; p = 0.95; r < 0.01; n = 49). There was no statistically significant relationship with the OECD status of the respondent and amounts awarded (F (1,45) < 0.001; p = 0.99; r = < 0.01; n = 49). There was also no statistically significant relationship between having a presiding arbitrator from an OECD state and amounts awarded (F (1,45) = 0.849; p = 0.36; r = 0.14; n = 49). For the main effect of a presiding arbitrator's OECD status, with a small effect, low power of 0.20 and an 80% risk of a Type II error, replication is warranted before making conclusions regarding the current population. For the other two effects with less than small effect sizes (power was less than 0.20), the failure to identify a statistically significant difference may not purely be a function of power. *See supra* note 159.

[172] For the inverse transformations on CPI-adjusted data, there was no statistically significant interaction with OECD status of the respondent, OECD status of the presiding arbitrator and amounts awarded (F (1,45) = 0.329; p = 0.57; r = 0.09; n = 49). There was no statistically significant relationship with the OECD status of the respondent and amounts awarded (F (1,45) = 0.102; p = 0.75; r = 0.05; n = 49). There was also no statistically significant relationship between having a presiding arbitrator from an OECD state and amounts awarded (F (1,45) = 1.468; p = 0.23; r = 0.18; n = 49). For the main effect of a presiding arbitrator's OECD status, with a small effect, low power of 0.20-0.30 and a 70-80% risk of a Type II error, replication is warranted before making conclusions regarding the current population. For the other two effects with less than small effect sizes (power was less than 0.20), the failure to identify a statistically significant difference may not purely be a function of power. *See supra* note 159.

We then repeated our work on the OECD-based models using the GDP-adjusted values. Again, to replicate the *Harvard International* article using a measure with decreased skewness, we selected (1) raw GDP-adjusted, (2) winsorized GDP-adjusted, (3) log GDP-adjusted, and (4) inverse GDP-adjusted data for analysis.[173] As with the CPI-adjusted values discussed above, when the analyses of the GDP-adjusted amounts awarded failed to reflect a statistically significant relationship between OECD-based development variables and outcome. First, using GDP-adjusted log transformations, there was no interaction among the OECD status of the respondent state, the OECD status of the presiding arbitrator and amounts awarded (F (1,45) = 0.067; p = 0.78; r = 0.04; n = 49).[174] Second, there was no statistically significant relationship between being an OECD respondent and amounts awarded (F (1,45) = 0.028; p = 0.87; r = 0.05; n = 49).[175] Third, there was no statistically significant relationship between having a presiding arbitrator from an OECD country and the amounts awarded (F (1,45) = 1.304; p = 0.26; r = 0.17; n = 49).[176] As with the CPI-adjusted analyses, when adjusting for GDP, irrespective of whether we used raw data,[177] winsorized data[178] or inverse transformations,[179] the models failed to reveal any

[173] The skewness of the GDP-adjusted amounts awarded was as follows: (1) raw = 5.493, (2) trimmed = 2.236, (3) winsorized = 1.421; (4) square root = 3.723, (5) log = 0.501, and (6) inverse = -0.404.

[174] The power of the GDP adjusted interaction was less than 0.20. Given the non-significant results and less than small effect size, the results may not be a function of power. *See supra* note 159.

[175] The power of the main effect of the respondent's OECD status using GDP adjusted amounts was less than 0.20. Given the non-significant results and less than small effect size it is not clear that the failure to identify a relationship is a function of power. *See supra* note 159.

[176] As with the CPI adjusted amounts, for the GDP adjusted log, the power of the main effect of the presiding arbitrator's OECD status was 0.20. Given the 80% risk of a Type II error, replication is warranted before making strong inferences beyond the pre-2007 population.

[177] For the raw GDP adjusted data, there was no statistically significant interaction with OECD status of the respondent, OECD status of the presiding arbitrator and amounts awarded (F (1,45) = 1.318; p = 0.26; r = 0.17; n = 49). There was no statistically significant relationship with the OECD status of the respondent and amounts awarded (F (1,45) < 0.001; p = 0.99; r = < 0.01; n = 49). There was also no statistically significant relationship between having a presiding arbitrator from an OECD state and amounts awarded (F (1,45) = 0.001; p = 0.97; r = 0.01; n = 49). For the interaction, with a small effect, low power of 0.20 and an 80% risk of a Type II error, replication is warranted before making definitive conclusions. For the two main effects with less than small effect sizes (power less than 0.20), the results may not be attributable to power problems. *See supra* note 159.

[178] For the winsorized GDP adjusted data, there was no statistically significant interaction with OECD status of the respondent, OECD status of the presiding arbitrator and amounts awarded (F (1,45) = 0.009; p = 0.92; r < 0.01; n = 49). There was no statistically significant relationship with the OECD status of the respondent and amounts awarded (F (1,45) = 0.002; p = 0.96; r = 0.01; n = 49). There was also no statistically significant relationship between

meaningful difference among the OCED-based variables and amounts awarded.

B. World Bank Based Models: CPI and GDP Adjusted Amounts Awarded

We then replicated the World Bank-based models using the CPI-adjusted values using the four CPI-adjusted variables described in Part A. As with the OECD-based model, none of the World Bank analyses revealed any statistically significant relationships between the World-Bank based variables and the CPI-adjusted amounts awarded. First, using the log transformation for CPI-adjusted amounts, there was no interaction among the World Bank status of the respondent state, the World Bank status of the presiding arbitrator and amounts awarded (F (3,41) = 0.711; p = 0.55; r = 0.13; n = 49).[180] Second, using log transformation, there was no statistically significant relationship between a respondent's World Bank classification and amounts awarded (F (3,41) = 1.108; p = 0.38; r = 0.16; n = 49).[181] Third, there was no statistically significant relationship between having a presiding arbitrator from a High or Middle Income country and amounts awarded (F (1,41) = 0.087; p = 0.77; r = 0.04; n = 49).[182] The World Bank results from the raw CPI-adjusted data were

---

having a presiding arbitrator from an OECD state and amounts awarded (F (1,45) = 0.835; p = 0.37; r = 0.14; n = 49). For the main effect of a presiding arbitrator's OECD status, with a small effect, low power of 0.20 and an 80% risk of a Type II error, replication is warranted before making conclusions regarding the current population. For the other two effects with less than small effect sizes (power was less than 0.20), the failure to identify a statistically significant difference may not purely be a function of power. *See supra* note 159.

[179] For the inverse transformations on GDP-adjusted data, there was no statistically significant interaction with OECD status of the respondent, OECD status of the presiding arbitrator and amounts awarded (F (1,45) = 0.329; p = 0.57; r = 0.09; n = 49). There was no statistically significant relationship with the OECD status of the respondent and amounts awarded (F (1,45) = 0.102; p = 0.75; r = 0.05; n = 49). There was also no statistically significant relationship between having a presiding arbitrator from an OECD state and amounts awarded (F (1,45) = 1.468; p = 0.23; r = 0.18; n = 49). For the main effect of a presiding arbitrator's OECD status, with a small effect, low power of between 0.20-0.30 and a 70-80% risk of a Type II error, replication is warranted before making conclusions regarding the current population. For the two less than small effects (power less than 0.20), the failure to identify a statistically significant difference may not purely be a function of power. *See supra* note 159.

[180] With the small effect (r = 0.13), low power of 0.20 and an 80% risk of a Type II error, replication is warranted before making conclusions regarding the current population.

[181] With the small effect (r = 0.16), low power of 0.20 and an 80% risk of a Type II error, replication is warranted before making conclusions regarding the current population.

[182] The power of the main effect of presiding arbitrators World Bank classification was less than 0.20. Given the non-significant results and the less than small effect size, the failure to identify a statistically significant difference is necessarily a function of power. *See supra* note 159.

similar.[183] Likewise, the results from the CPI-adjusted winsorized[184] data and inverse transformations[185] failed to reveal any statistically significant results in the World Bank-based models.

Finally, we replicated the World Bank-based models using the four GDP-adjusted values described in Part A. Like the other analyses, the results failed to identify a statistically significant relationship with World Bank-based development variables and the GDP-adjusted amounts awarded. First, using the log transformation for GDP-adjusted amounts, there was no interaction among the World Bank status of the respondent state, the World Bank status of the presiding arbitrator and amounts awarded (F (3,41) = 0.704; p = 0.56; r = 0.22; n = 49).[186] Second, using log transformations, there was no

[183] There was no statistically significant interaction with World Bank status of the respondent, World Bank status of the presiding arbitrator and amounts awarded (F (3,41) = 0.688; p = 0.56; r = 0.22; n = 49). There was no statistically significant relationship with the World Bank status of the respondent and amounts awarded (F (3,41) = 0.173; p = 0.91; r = 0.11; n = 49). There was also no statistically significant relationship between a presiding arbitrator's World Bank classification and amounts awarded (F (1,41) = 0.060; p = 0.81; r = 0.04; n = 49). For the main effect of the presiding arbitrator, with less than small effect size and low power (less than 0.20), the failure to identify a significant difference may not purely be a function of power *See supra* note 159. The interaction, however, was underpowered (less than 0.20 for small effect); and the main effect of the respondent's development status exhibited low power (0.30). The resulting 70-80% risk of a Type II error for those models indicates replication is warranted before making conclusions beyond the pre-2007 population.

[184] For the winsorized data, there was no statistically significant interaction with World Bank status of the respondent, World Bank status of the presiding arbitrator and amounts awarded (F (3,41) = 0.518; p = 0.67; r = 0.19; n = 49). There was no statistically significant relationship with the World Bank status of the respondent and amounts awarded (F (3,41) = 1.246; p = 0.31; r = 0.29; n = 49). There was also no statistically significant relationship between having a High or Middle Income presiding arbitrator and amounts awarded (F (1,41) = 0.392; p = 0.53; r = 0.10; n = 49). As all of these models had small or a medium effect, the power of the analyses was less than 0.20 (r = 0.10), 0.30 (r = 0.19) and in the range of 0.50-0.60 (r = 0.29). Given the risk of a Type II error, replication is warranted before making conclusions regarding the current population.

[185] For the inverse transformations on CPI-adjusted data, there was no statistically significant interaction with World Bank of the respondent, World Bank status of the presiding arbitrator and amounts awarded (F (3,41) = 0.751; p = 0.53; r = 0.23; n = 49). There was no statistically significant relationship with the World Bank status of the respondent and amounts awarded (F (3,41) = 0.990; p = 0.41; r = 0.26; n = 49). There was also no statistically significant relationship between a presiding arbitrator's World Bank status and amounts awarded (F (1,41) = 0.060; p = 0.81; r = 0.04; n = 49). For the main effect of a presiding arbitrator's World Bank status, with a less than small effect and non-significant result, the failure to identify a relationship may not just be a function of power. *See supra* note 159. As the two other effects, however, have low power (0.30-0.40), the 60-70% risk of a Type II error warrants replication before making conclusions beyond the current dataset.

[186] With the small-medium effect (r = 0.22), low power of 0.30 and a 70% risk of a Type II error, replication is warranted before making conclusions regarding the current population.

statistically significant relationship between a respondent's World Bank classification and amounts awarded (F (3,41) = 1.105; p = 0.36; r = 0.27; n = 49).[187] Third, the last log transformation of GDP-adjusted data identified no statistically significant relationship between having a presiding arbitrator from a High or Middle Income country and amounts awarded (F (1,41) = 0.090; p = 0.76; r = 0.09; n = 49).[188] As with the CPI-adjusted analyses, when adjusting for GDP, irrespective of whether we used raw data,[189] winsorized data[190] or inverse transformations,[191] the models failed to reveal any meaningful difference among the World Bank-based variables and amounts awarded.

[187] With a nearly medium effect (r = 0.27), power of 0.40 and a 60% risk of a Type II error, replication is warranted before making conclusions regarding the current population.

[188] The power of the main effect of presiding arbitrators World Bank classification was less than 0.20. Given the non-significant results and the less than small effect size, the failure to identify a statistically significant difference may not purely be a function of power. *See supra* note 159.

[189] For the raw GDP adjusted data, there was no statistically significant interaction with World Bank status of the respondent, World Bank status of the presiding arbitrator and amounts awarded (F (3,41) = 0.668; p = 0.58; r = 0.22; n = 49). There was no statistically significant relationship with the World Bank status of the respondent and amounts awarded (F (3,41) = 0.160; p = 0.92; r = 0.11; n = 49). There was also no statistically significant relationship between having a presiding arbitrator from a High or Middle Income state and amounts awarded (F (1,41) = 0.052; p = 0.82; r = 0.04; n = 49). For the less than small effect of the presiding arbitrator's World Bank status, the results may not be attributable to a power problem. However, given the low power for the interaction (0.30) and main effect of the respondent's World Bank status (less than 0.20) given the risk of a Type II error, replication is warranted before making conclusions regarding the current population.

[190] For the winsorized GDP-adjusted data, there was no statistically significant interaction with World Bank status of the respondent, World Bank status of the presiding arbitrator and amounts awarded (F (3,41) = 0.516; p = 0.67; r = 0.19; n = 49). There was no statistically significant relationship with the World Bank status of the respondent and amounts awarded (F (3,41) = 1.233; p = 0.31; r = 0.29; n = 49). There was also no statistically significant relationship between having a presiding arbitrator from a High or Middle Income state and amounts awarded (F (1,41) = 0.402; p = 0.53; r = 0.10; n = 49). As all the effects ranged from small to medium, the low power ranged from less than 0.20 to 0.60. Given the 40-80% risk of a Type II error, replication is warranted before making conclusions regarding the current population.

[191] For the inverse transformations on GDP-adjusted data, there was no statistically significant interaction with World Bank of the respondent, World Bank status of the presiding arbitrator and amounts awarded (F (3,41) = 0.751; p = 0.53; r = 0.23; n = 49). There was no statistically significant relationship with the World Bank status of the respondent and amounts awarded (F (3,41) = 0.990; p = 0.41; r = 0.26; n = 49). There was also no statistically significant relationship between a presiding arbitrator's World Bank status and amounts awarded (F (1,41) = 0.060; p = 0.81; r = 0.04; n = 49). For the main effect of a presiding arbitrator's World Bank status, with a less than small effect and non-significant result, the failure to identify a relationship may not necessarily reflect a power problem. *See supra* note 159. As the two other effects, however, have low power (0.30-0.40), the 60-70% risk of a Type II error warrants replication before making conclusions beyond the current dataset.