OSGOODE HALL LAW SCHOOL | YORK UNIVERSITÉ UNIVERSITY

Osgoode Hall Law School of York University
Osgoode Digital Commons

All Papers

Research Papers, Working Papers, Conference Papers

2010

Fairness and Independence in Investment Arbitration: A Critique of Susan Franck's "Development and Outcomes of Investment Treaty Arbitration"

Gus Van Harten
Osgoode Hall Law School of York University, gvanharten@osgoode.yorku.ca

Follow this and additional works at: http://digitalcommons.osgoode.yorku.ca/all_papers

Repository Citation
Van Harten, Gus, "Fairness and Independence in Investment Arbitration: A Critique of Susan Franck's "Development and Outcomes of Investment Treaty Arbitration"" (2010). *All Papers*. Paper 30.
http://digitalcommons.osgoode.yorku.ca/all_papers/30

This Working Paper is brought to you for free and open access by the Research Papers, Working Papers, Conference Papers at Osgoode Digital Commons. It has been accepted for inclusion in All Papers by an authorized administrator of Osgoode Digital Commons.

Fairness and independence in investment arbitration: A critique of Susan Franck's "Development and Outcomes of Investment Treaty Arbitration"

Gus Van Harten*

In this article, I review a prominent study, Franck (2009),[1] that used quantitative research tools to evaluate possible bias in investment arbitration. The study has been cited in policy and academic discussions in order to allay concerns about the fairness and independence of investment arbitration. I argue here that findings and conclusions reached in the study were exaggerated or unsupported by the results of the study.

In summary:

- Franck's study suffered from a lack of data, thus undermining its reliability. The lack of data led to a 40 to 80% of error across most of Franck's results. Franck identified this high risk of error in a series of footnotes in the study but did not mention it alongside prominent claims that "development status does not have a statistically significant relationship with outcome" and that "outcome is not associated with arbitrator or respondent development status". These claims were inaccurate in light of the high risk of error. The most clearly-supported conclusion of the study, although not identified by Franck, was that there was insufficient data to test the hypothesis with an acceptable level of reliability.

- The study lacked validity in its measurement of development status. By using OECD membership as a proxy for developed-country status, Franck treated Mexico and former East Bloc countries as developed countries. In her results, 8 of the 18 cases (44%) classified by Franck as against developed countries were cases against Mexico and 3 of these 18 cases (17%) were against the Czech Republic and the Slovak Republic. Although known to her prior to publication, Franck did not disclose this aspect of her results and explain why she considered it appropriate to treat Mexico and former East Bloc countries as developed countries.

- The study is relevant to expectations of actual bias, observable at a systemic level, arising from aspects of the nationality of arbitrators and the identity of respondent states. It is of little or no relevance to questions of actual bias in specific cases and of perceptions of bias arising from institutional factors. Regardless of its results, the study could not provide a "powerful narrative" for or against the procedural integrity of the system, as was claimed by Franck. Likewise, the statement that the study "suggests that the investment treaty arbitration system, as a whole, functions fairly and that the eradication or radical overhaul of the arbitration process is

* Associate Professor, Osgoode Hall Law School.

[1] S.D. Franck, "Development and Outcomes of Investment Treaty Arbitration" (2009) 50 *Harvard International Law Journal* 435. Further methodological background to this study was provided in S.D. Franck, "Empirically Evaluating Claims About Investment Treaty Arbitration" (2007) 80 *North Carolina Law Review* 1.

> unnecessary"[2] was inappropriate in light of the lack of validity and reliability of the study. Much more information involving a wide range of factors would be required to contemplate such claims.

This critique does not lead to the conclusion that there is evidence of actual bias in investment arbitration. The main lessons are that (1) there is far too little information to draw reliable conclusions either way based on quantitative research and (2) it is important to present empirical research with care and accuracy in order not to mislead policy-makers and academic commentators alike.

More broadly, the key institutional concern in investment arbitration is to ensure that reasonable perceptions of bias are addressed. The best way to do this is by introducing well-known institutional safeguards of independence in the process, not by seeking empirical proof or dis-proof of actual bias.

Some concerns about perceived bias

Investment treaty arbitration is unlike other forms of arbitration and international adjudication. It empowers arbitrators to make final decisions on public law and on important policy concerns. It raises issues of independence and impartiality that generally do not arise in other forms of arbitration.

In other contexts, both domestic and international, public law is decided finally by judges whose independence from state and private power is protected by institutional safeguards, including secure tenure, bars on outside remuneration, and an objective method of case assignment. The absence of these safeguards in investment treaty arbitration raises a reasonable perception in all cases that inappropriate factors have influenced a decision or award.

One set of possible influences arises from the financial and career interests of arbitrators who lack secure tenure and who engage in remunerative activities outside of the adjudicative role. Another arises from the potential influence of arbitral institutions and of private actors in the arbitration industry. Of course, the presence of these concerns does not explain fully the expected behaviour of arbitrators. One hopes and trusts that other considerations, values of fairness and integrity, will drive decisions.

The problem is that no one—other than the individual decision-maker—can know whether inappropriate factors have come into play. For this reason, the actual behaviour of arbitrators is not the sole concern. As important is the role of institutional safeguards in addressing reasonable perceptions of bias.

A final point is that openness is integral to independence and impartiality. Without openness, it is not possible to verify the fairness and integrity of a decision-making process. All empirical research on investment arbitration confronts this problem. At

[2] Franck (2009), above note 1, p. 435.

present, in some cases, we do not know who made the decisions, what decisions were made, and what policy concerns arose.[3] This fuels concerns about unfairness for the very reason that the process is being kept secret.

Comment on Franck (2009)

Empirical research can contribute, alongside deductive reasoning and doctrinal analysis, to the development of knowledge about investment law and its institutions. Yet empirical researchers must be clear about the questions they are examining and about the limitations and qualifications of their conclusions. These may seem like obvious points, but they are important to stress.

A number of studies have reported on outcomes in known investment arbitrations. Some commentators have relied on such data to advance claims about the actual performance of arbitrators. Many of the studies to date, however, do not examine specific hypotheses of bias or position the study in terms of literature on institutional aspects of adjudicative independence. Existing studies also face serious methodological constraints and depend on assumptions that heavily qualify results.[4] As such, one should be very cautious about using such studies to draw conclusions about the actual behaviour of arbitrators.

A few studies have sought to analyze specific hypotheses of possible bias on the part of investment arbitrators. They have focused on actual bias (usually at a systemic level) as opposed to reasons for perceived bias. The study by Franck is probably the most prominent.[5]

Outline of the study

In the study, Prof. Franck examined hypotheses arising from individual factors that could generate actual bias in investment arbitration. These factors involved possible arbitrator prejudices tied to their nationality and/ or to characteristics of the respondent

This is especially true for arbitrations conducted in private forums such as the International Chamber of Commerce.

[4] For example, data on outcomes is typically drawn from ICSID, because of the relative openness of ICSID proceedings, and may say little or nothing about other forums. Also, data on outcomes as a measure of actual performance is open to a range of alternative explanations, such as variations in the strength of parties' claims, diversity of fact situations, possible inflation by claimants of amounts claimed, procedural variations among forums, varying experience levels and incentives among arbitrators, and varying political influences of states and private actors. Existing data on outcomes also does not capture aspects of tribunal decisions – such interpretations of the law – that may reflect bias independently of outcome. It is also very difficult, if not impossible, to identify the "appropriate" or "fair" spread of outcomes against which actual outcomes are to be measured. Lastly, cumulative data on outcomes does not explain whether the outcome in any particular case was influenced by inappropriate factors. As such, there will always remain a basis for perceived pro-investor or pro-state bias in specific instances based on inadequacies of the institutional structure. This is the concern that institutional safeguards of independence are meant to address.

[5] Franck (2009), above note 1.

state. Both of these factors were grouped according to the "development status" of arbitrators based on their nationality and respondent countries, and then compared to outcomes in specific cases.

Franck's hypothesis was that development status would not affect outcome and "that arbitrators can make decisions neutrally on the basis of the facts and law".[6] The study design tested only the first element of this hypothesis, but was evidently intended to provide a basis for comment on the second.

To test the hypothesis, Franck analyzed outcomes in 52 treaty cases with publicly-available information. She applied two metrics to classify the development status of the presiding arbitrators and respondent states. The first metric was OECD membership. Arbitrators and countries were treated as "developed" if they (or their countries of nationality) were members of the OECD and as "developing" if they were non-members. Second, the study classified arbitrators and countries based on the World Bank income classification system.[7]

Based on this analysis, the study did not find significant variations between outcomes linked to the development status of presiding arbitrators and respondent countries. In turn, Franck drew some bold conclusions about the integrity and fairness of investment arbitration.

Franck's outline of her methodology is commendable for its clarity and transparency. However, the study has limitations and, in some cases, important flaws. Most important is the extent to which Franck over-stated or mis-stated key conclusions. The following are examples, drawn from the main text, conclusion, or abstract of the study.

- In the main text regarding the comparison of development status to win/ loss outcomes, it is stated that the study "offers a powerful narrative that there is procedural integrity in investment arbitration".[8]

- In the main text regarding the comparison of development status and amount-of-damages outcomes, it is stated that the "lack of a main effect for a respondent's development status stands in sharp contrast to the assertions

[6] Franck (2009), above note 1, p. 454.
[7] The World Bank income classification system divides countries, based on Gross National Income per capita, into four categories: high income, upper middle income, lower middle income, and low income. In classifying outcomes, Franck characterized as a win for the respondent state an award of no damages against the state. She treated as a loss an award of any amount in excess of zero damages against the state. Second, Franck evaluated outcome in terms of the amounts of money awarded in cases where the claimant was successful. Due to limitations in the available data, Franck's sample was reduced to outcomes in 49 cases for the OECD metric and 47 or 49 cases (depending on whether a win/ loss or total damages outcome was being measured) for the World Bank metric.
[8] Franck (2009), above note 1, p. 464.

that investment treaty arbitration unfairly privileges the developed world or improperly harms the developing world."[9]

- In the study's conclusion it is stated:

  "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 affects such outcomes. It suggests that major structural overhaul may not be necessary because it is not clear that arbitration is inherently predisposed towards particular outcomes."[10]

- In the study's abstract it is stated:

  "The results demonstrate that, at the macro level, development status does not have a statistically significant relationship with outcome. This suggests that the investment treaty arbitration system, as a whole, functions fairly and that the eradication or radical overhaul of the arbitration process is unnecessary."[11]

For the reasons outlined below, these statements, to varying degrees, are exaggerated or misplaced.

Limitations of the OECD metric

An initial limitation of the study relates to the use of the OECD metric. The equation of OECD membership with developed status makes it likely that OECD countries that are reasonably classified as developing or transition countries (Mexico, Turkey, and the former East Bloc OECD countries) were classified as "developed". This raises the prospect of a misclassification of the development status of arbitrators or countries in the data.

This limitation would not be serious if it were communicated clearly and transparently by the researcher. However, Franck did not identify this limitation of the OECD metric in her study. She also did not attempt to indicate whether and how the findings might vary if the data was broken down in order to separate developed OECD countries from developing/ transition OECD countries.

I examined Franck's data in order to determine whether accounting for this aspect of the OECD metric affected the results. This was done by distinguishing Mexico, South

[9] Franck (2009), above note 1, p. 470.
[10] Franck (2009), above note 1, p. 487.
[11] Franck (2009), above note 1, p. 435.

Korea, and the former East Bloc countries from other OECD members.[12] The review confirmed that this limitation of the OECD metric affected significantly the study's results. In particular:

- In the 49 cases reviewed by Franck using the OECD metric, all 36 of the presiding arbitrators who were nationals of an OECD country were nationals of a "developed" OECD country. This included arbitrators from 13 countries: United Kingdom (6 cases), Sweden (5), Australia (4), Germany (4), USA (4), Switzerland (3), Canada (2), France (2), Spain (2), Denmark (1), Greece (1), Italy (1), and the Netherlands (1).

- Of the 36 tribunals with a presiding arbitrator who was a national of a developed OECD country, 10 decided cases against a developing or transition OECD country (including 7 cases against Mexico, 2 against the Czech Republic, and 1 against Slovakia).

- Of the 13 tribunals with a presiding arbitrator who was not a national of an OECD country, 1 decided a case against a developing OECD country (Mexico).

- Accounting for the heterogeneity of OECD membership, then, 10 cases should have been classified as developed-to-developing or developed-to-transition arbitrations rather than as developed-to-developed arbitrations. Also, 1 case should have been classified as a developing-to-developing rather than a developing-to-developed arbitration.

Thus, in Franck's study, 11 of the 49 cases were arguably mis-classified. By accounting for the heterogeneity of OECD membership, the ratio of developed to developing respondent countries in Franck's data dropped dramatically from a ratio of 18 to 31 to a ratio of 7 to 42.

The limitations of the OECD metric are not the critical point here. The key issue is the lack of transparency about a methodological limitation that undermines the validity the study. As outlined, use of the OECD metric arguably led to a misclassification of 22% of cases. But the study does not highlight this concern to the reader or attempt to examine how it affected the results.[13]

[12] I examined OECD membership in 2009, which was the year of publication of Franck's study. Thus, Chile, Israel, and Slovenia – which joined the OECD in 2010 – were characterized as non-OECD members for purposes of this review of Franck's data.

[13] Note that this criticism of the OECD metric was conveyed to Professor Franck prior to publication of her 2009 study, above note 1. Other criticisms laid out in this article were not conveyed until after publication of the study.

Collapsing of the World Bank metric

A second limitation involved the World Bank income classification metric that was used by Franck to measure development status. This limitation arose from a lack of data.

Franck sought to compare income levels of the countries of nationality of presiding arbitrators to the income levels of respondent states. As such, there were 16 boxes in which the study required data in order to test her hypotheses using the 4-level World Bank metric. However, Franck did not have enough data to do this. First, there were no cases at all decided by presiding arbitrators from low income countries; thus, four of the 16 boxes in Franck's analysis contained no data. Second, there was only one case that was decided by a presiding arbitrator from a lower middle income country.

In response to this lack of data, Franck collapsed the 16-box grid into an 8-box grid.[14] In turn, all but one of the cases decided by a "developed country" presiding arbitrator involved arbitrators from upper-middle income countries. Franck was transparent about the need to collapse the World Bank metric. This allows the reader to see that the World Bank metric was frustrated by a lack of data.

On the other hand, the limitations of the World Bank and the OECD metrics highlight that the study tested only very narrow aspects of questions about possible bias, let alone fairness and independence, in investment arbitration. Specifically, the study relied on a limited data pool to test whether there was a connection between nationality groupings of arbitrators or countries and particular outcomes. Regardless of its results, the study could not provide a "powerful narrative" either for or against the procedural integrity of the system or support conclusions on whether the system "functions fairly". Far more information involving a wide range of factors would be required to contemplate such claims.

Lack of data and corresponding risk of error

A major limitation of the study is common in empirical work on investment arbitration. This is the lack of data, which in turn raises concerns about reliability.

The lack of data in this case frustrated the drawing of reliable conclusion to support or refute Franck's hypothesis that development status would not affect outcome.[15] Indeed, the only clearly-supported conclusion of Franck's study was that there was insufficient data to test her hypothesis with an acceptable level of reliability. This lack of reliability affected all four of the metrics employed by the study.

A claim of statistical significance about a hypothesized connection (or lack of connection) between variables requires sufficient data to remove any significant risk that the apparent relationships are explained by chance. Franck calculated that the number of

[14] Franck (2009), above note 1, p. 459.
[15] Franck (2009), above note 1, p. 454.

awards needed to generate statistically significant findings (on her standard, findings that carried a 20% chance of error). Depending on the metric and the effect size of the results, between 382 and 781 awards were required for most of Franck's comparisons.[16] Yet the available sample sizes were between 47 and 49 awards.

This led to a 40 to 80% chance of error (so-called "Type II error", referring to the risk of a false negative) in Franck's results. As a result, there was insufficient evidence of either the presence or absence of a statistically significant connection between development status and outcome. Findings or conclusions beyond this, based on the results, lack reliability due to the high risk of error.

It was rigorous and transparent for Franck to provide these calculations on the risk of error and to indicate the corresponding limitations. But these issues were reported only in a series of footnotes and, in some instances, tangentially in the main text.[17] Importantly, Franck did not mention the high risk of error underlying the bold claims reproduced above.

More fundamentally, the study relied on the risk of error associated with a hypothesized connection between development status and outcome in order to convey that there was in fact no connection between development status and outcome.[18] The latter does not necessarily follow from the former. The more appropriate conclusion to draw was that there was insufficient data to test the hypothesis that development status would not affect outcome.

For these reasons, Franck's results do not establish that "development status does not have a statistically significant relationship with outcome", for example, as was claimed. Development status may or may not affect outcome; based on Franck's study, we do not know with a reasonable degree of reliability. Likewise, the high risk of error should have precluded Franck from reporting that "outcome is not associated with arbitrator or respondent development status".

Failure to account for alternative explanations

In empirical research, there is typically a range of alternative explanations for results. It is important for a researcher to convey clearly that such alternatives exist and to avoid undue emphasis on one or a small number of possible explanations.

However, Franck emphasized only a limited set of explanations for her results—such as the prospect that the system functions fairly—while neglecting others. One alternative explanation, for example, is that arbitrators do not make decisions based on nationality, but rather based on their membership in a common culture or industry of arbitrators. Another is that the facts of cases involving some classes of countries are more

[16] Franck (2009), above note 1, p. 461-70.
[17] Franck (2009), above note 1, p. 461-70.
[18] Franck (2009), above note 1, p. 460-62.

favourable to investors than the facts of cases involving others countries, and that factual differences will lead to variations in outcome.

Even assuming that her results were reliable, it was an over-reach for Franck, in light of alternative explanations, to draw conclusions on the "integrity of arbitration", on whether the system "functions fairly", or on whether there is need for a "major structural overhaul". A wide range of institutional and individual concerns about potential bias were simply not tested by the study.

Conclusions

These issues and concerns do not raise problems with the empirical method itself, but rather with the way in which the method was employed in Franck's study. Many of the limitations of the study could have been avoided with greater transparency and caution on the part of the researcher in the statement of conclusions.

The fundamental problem is the way in which Franck constituted and presented her conclusions. The key problems are:

- the claim that a lack of reliable evidence of a connection between nationality and outcome demonstrated the absence of such a connection,

- the failure to make clear that the lack of data precluded the study's hypothesis from being tested reliably, and

- the failure to highlight alternative explanations for the results in statements of the conclusions.

This discussion highlights how important it is for a researcher to present findings and conclusions accurately and with care. Otherwise, there is a danger that policy makers will take up a study for purposes that the research does not support, as has happened in the case of Franck's work.[19]

Most importantly, it would be a mistake to rely on quantitative methods to address perceived bias in adjudication. Absent an admission by the decision-maker, it is not possible to show definitively whether inappropriate factors have affected a particular decision or award. This explains why it is critical for the institutional structure of adjudication to allay concerns about the financial or career interests of adjudicators, regardless of whether or not these factors are affecting actual decisions. The priority is not to seek definitive proof or dis-proof of actual bias but rather to anticipate and

[19] See the separate statement by a member of the U.S. Model BIT Review committee, Mark Kantor, who cited Franck's conclusion on development status and outcome in support of his position against reform of the U.S. Model BIT. United States, *Report of the Advisory Committee on International Economic Policy Regarding the Model Bilateral Investment Treaty* (30 September 2009), Annex B (Particular Viewpoints of Subcommittee Members), Statement of Mark Kantor, p. 28-9.

address the uncertainties that give rise to reasonably perceived bias. At the institutional level, this calls for the incorporation of well-known safeguards of judicial independence in order to support public confidence in the fairness of adjudication.