The European Journal of International Law Vol. 31 no. 2

Challenges and Pitfalls in Research on Compliance with the 'Views' of UN Human Rights Treaty Bodies: A Reply to Vera Shikhelman

Andreas J. Ullmann* and Andreas von Staden**

Abstract

Vera Shikhelman's recent article on the implementation of the views of the United Nations Human Rights Committee (HRC) took a valuable first step towards addressing the question why states do, or do not, comply with adverse treaty body views. In this contribution, we contend that parts of the chosen theoretical and methodological approach are somewhat problematic, however, and ultimately weaken the overall strength and reliability of Shikhelman's findings. Theoretically, we question whether hypotheses developed in the context of studying compliance with international law and legally binding court judgments can be transferred to legally non-binding views without adjusting for potentially consequential differences in their legal status. Methodologically, we note certain issues concerning the data generated by the HRC's follow-up procedure and its use in Shikhelman's analysis, and suggest that statistical methods that take into consideration the time dimension of implementation processes, notably survival analysis, yield analytically more convincing causal inferences. We provide illustrative results of such a methodologically revised approach to examining compliance with adverse HRC views that reveal more fine-grained insights into the temporally unfolding processes of implementing such decisions.

* Universität Hamburg, Germany. Email: andreas.ullmann@uni-hamburg.de.
** Universität Hamburg, Germany. Email: andreas.vonstaden@uni-hamburg.de.

1 Introduction

Vera Shikhelman's article on the implementation of the 'views' of the United Nations (UN) Human Rights Committee (HRC) is a welcome addition to the literature on compliance with the output of international monitoring and dispute settlement bodies in the human rights domain.[1] Whereas social scientists have explored the compliance records of the two regional human rights courts in the Americas and in Europe for some time now,[2] research on the global human rights treaties has been dominated by questions of why states commit to such treaties in the first place,[3] especially when they are in obvious violation of the treaty's terms at the time of ratification,[4] and whether treaty ratification as such has an impact on states' human rights performance,[5] usually measured using high-level indicators such as the Political Terror Scale and the Freedom House Index. State acceptance of optional monitoring mechanisms, such as the individual communications procedures, has also received some attention,[6] but the question of the extent to which states comply with adverse decisions produced

1 Shikhelman, 'Implementing Decisions of International Human Rights Institutions: Evidence from the United Nations Human Rights Committee', 30 *European Journal of International Law* (*EJIL*) (2019) 753.

2 Hawkins and Jacoby, 'Partial Compliance: A Comparison of the European and Inter-American Courts of Human Rights', 6 *Journal of International Law and International Relations* (2010) 35; D. Anagnostou (ed.), *The European Court of Human Rights: Implementing Strasbourg's Judgments on Domestic Policy* (2013); C. Hillebrecht, *Domestic Politics and International Human Rights Tribunals: The Problem of Compliance* (2014); A. von Staden, *Strategies of Compliance with the European Court of Human Rights*: *Rational Choice within Normative Constraints* (2018).

3 Hathaway, 'The Cost of Commitment', 55 *Stanford Law Review* (2003) 1821; Hathaway, 'Why Do Countries Commit to Human Rights Treaties?', 51 *Journal of Conflict Resolution* (2007) 588; Cole, 'Sovereignty Relinquished? Explaining Commitment to the International Human Rights Covenants, 1966–1999', 70 *American Sociological Review* (2005) 472; Wotipka and Tsutsui, 'Global Human Rights and State Sovereignty: State Ratification of International Human Rights Treaties, 1965–2001', 23 *Sociological Forum* (2008) 724; B. Simmons, *Mobilizing for Human Rights: International Law in Domestic Politics* (2009), ch. 3; Sandholtz, 'Domestic Law and Human Rights Treaty Commitments: The Convention against Torture', 16 *Journal of Human Rights* (2017) 25.

4 Goodliffe and Hawkins, 'Explaining Commitment: States and the Convention against Torture', 68 *Journal of Politics* (2006) 358; Vreeland, 'Political Institutions and Human Rights: Why Dictatorships Enter into the United Nations Convention against Torture', 62 *International Organizations* (*IO*) (2008) 65; Powell and Staton, 'Domestic Judicial Institutions and Human Rights Treaty Violation', 53 *International Studies Quarterly* (*ISQ*) (2009) 149; Hollyer and Rosendorff, 'Why Do Authoritarian Regimes Sign the Convention against Torture? Signaling, Domestic Politics and Non-Compliance', 6 *Quarterly Journal of Political Science* (2011) 275; M.H. Hong, 'Crafting Reputation before Domestic and International Audiences: Autocratic Participation in the United Human Rights Institutions' (2016) (PhD thesis on file at University of Michigan), available at https://deepblue.lib.umich.edu/handle/2027.42/133208.

5 Keith, 'The United Nations International Covenant on Civil and Political Rights: Does It Make a Difference in Human Rights Behavior?', 36 *Journal of Peace Research* (1999) 95; Hathaway, 'Do Human Rights Treaties Make a Difference?', 111 *Yale Law Journal* (2002) 1935; Simmons, *supra* note 3; Conrad and Hencken Ritter, 'Treaties, Tenure, and Torture: The Conflicting Domestic Effects of International Law', 75 *Journal of Politics* (2013) 397.

6 Hathaway, *supra* note 3; Cole, *supra* note 3; H. Smith-Cannoy, *Insincere Commitments: Human Rights Treaties, Abusive States, and Citizen Activism* (2012); Hong, *supra* note 4; Hong and Uzonyi, 'Deeper Commitment to Human Rights Treaties: Signaling and Investment Risk Perception', 44 *International Interactions* (2018) 1040.

by these procedures, and why or why not, has so far been largely ignored, apart from some legal-descriptive[7] and macro-correlational analysis.[8] Shikhelman is thus correct in noting that '[t]here is very little literature that explores specific steps taken by states to implement decisions and recommendations of [non-judicial] international human rights institutions'.[9] Her work is to be commended for taking a first step towards remedying this lacuna and for opening wide the door for further research into why states do or do not comply with adverse treaty body views. Given the treaty bodies' rising output and the focus on the system as a whole as part of the 2020 treaty body review process,[10] such research will likely increase in amount.

That said, we perceive a number of issues in Shikhelman's article that threaten to weaken, if not undermine, her findings and which should be addressed and resolved in future research on compliance with treaty body views.[11] These issues concern theoretical expectations as to the factors driving compliance and non-compliance with legally non-binding views, her use of the available data and the study's methodological set-up and execution. In the following parts, we address each of these issues in turn.

2 Theoretical Considerations

Shikhelman's theoretical expectations are squarely grounded in existing research on compliance (and non-compliance) with international law generally, and in the decisions of international judicial institutions specifically, and the variables and indicators she uses to operationalize her hypotheses are familiar to readers of compliance studies. What is missing, though, is a theoretical discussion of the extent to which the underlying causal mechanisms that are identified as relevant in the law- and court-focused literature can be expected to work identically, or similarly, in the specific context of compliance with HRC views in light of their legally non-binding nature.[12] In other words, to what extent should we expect that differences in legal bindingness

7 Fox Principi, 'United Nations Individual Complaints Procedures: How Do States Comply?: A Categorized Study Based on 268 Cases of "Satisfactory" Implementation under the Follow-Up Procedure, Mainly Regarding the UN Human Rights Committee', 37 *Human Rights Law Journal* (*HRLJ*) (2017) 1; Fox-Principi, 'Internal Mechanisms to Implement U.N: Human Rights Decisions, Notably of the U.N. Human Rights Committee', 37 *HRLJ* (2017) 237.

8 Cole, 'Individuals v. States: The Correlates of Human Rights Committee Rulings, 1979–2007', 40 *Social Science Research* (2011) 985.

9 Shikhelman, *supra* note 1, at 754.

10 See generally https://www.ohchr.org/EN/HRBodies/HRTD/Pages/TBStrengthening.aspx.

11 For purposes of full disclosure, we note that we ourselves are currently engaged in just such a project, titled 'On the Causal (In)Significance of Legal Status: Assessing and Explaining Compliance with the "Views" of the UN Human Rights Treaty Bodies' and funded by the Deutsche Forschungsgemeinschaft (German Research Foundation), project no. 417704617.

12 Herndl, 'Zur Frage des rechtlichen Status der Entscheidungen eines Staatengemeinschaftsorgans: Die "Views" des Menschenrechtsausschusses', in K. Ginther *et al.* (eds), *Völkerrecht zwischen normativem Anspruch und politischer Realität: Festschrift für Karl Zemanek zum 65. Geburtstag* (1993) 205. For a so far singular domestic decision proclaiming the binding effect of the Committee on the Elimination of Discrimation against Women's views, see Kanetake, 'María de los Ángeles González Carreño v. Ministry of Justice, Judgment no. 1263/2018, Supreme Court of Spain, July 17, 2018', 113 *American Journal of International Law* (*AJIL*) (2019) 586.

will be consequential in the context of the hypotheses tested by Shikhelman that take their cues from the theories developed with respect to legally binding treaties and judgments? This is not a trivial theoretical issue as the recurrent debates about the appropriate characterization and legal significance of 'soft law'[13] and the different degrees of 'legalization'[14] and of compliance with such instruments (as well as other forms of influence and impact) indicate.[15]

The position one takes on the issue of the significance of the presence (or absence) of legal bindingness may affect the hypothesized direction of causal factors and/or the expected magnitude of their effects. Shikhelman, for example, refers to findings in the literature that 'states with a strong rule of law … are more likely to implement judgments',[16] but it does not necessarily follow that this has to hold with respect to legally non-binding pronouncements as well. Indeed, we could reasonably expect either positive or negative effects of the rule of law on compliance, the former fueled by the tendency of rule-of-law systems to provide effective remedies, the latter informed by the fact that the legally non-binding views of the HRC will often conflict with legally binding domestic decisions, which rule-of-law countries might be inclined to give priority over non-law.[17] Elsewhere, again well in line with the expectations and findings of the human rights literature focusing on non-governmental organization activities,[18] the existence of a '[s]trong civil society' should be reasonably associated with better implementation records,[19] but one of the causal pathways mentioned – assisting with litigation to implement remedies – depends crucially on whether the domestic legal

[13] See, e.g., Klabbers, 'The Redundancy of Soft Law', 65 *Nordic Journal of International Law* (*NJIL*) (1996) 167; Klabbers, 'The Undesirability of Soft Law', 67 *NJIL* (1998) 381; Raustiala, 'Form and Substance in International Agreements', 99 *AJIL* (2005) 581; Guzman and Meyer, 'Soft Law', in E. Kontorovich and F. Parisi (eds), *Economic Analysis of International Law* (2016) 123; Broude and Shereshevsky, 'Explaining the Practical Purchase of Soft Law: Competing and Complementary Behavioral Hypotheses', Hebrew University of Jerusalem Legal Research Paper 18-7 (2017), available at https://ssrn.com/abstract=3087179; P. Westerman *et al.* (eds), *Legal Validity and Soft Law* (2018).

[14] See Goldstein *et al.* (eds), *Legalization and World Politics* (2001); Finnemore and Toope, 'Alternatives to "Legalization": Richer Views of Law and Politics', 55 *IO* (2001) 743.

[15] See, e.g., D. Shelton (ed.), *Commitment and Compliance: The Role of Non-Binding Norms in the International Legal System* (2000); Druzin, 'Why Does Soft Law Have Any Power Anyway?', 7 *Asian Journal of International Law* (2016) 361; S. Lagoutte, T. Gammeltoft-Hansen and J. Cerone (eds), *Tracing the Roles of Soft Law in Human Rights* (2016).

[16] Shikhelman, *supra* note 1, at 759.

[17] For examples of such reasoning, see van Alebeek and Nollkaemper, 'The Legal Status of Decisions by Human Rights Treaty Bodies in National Law', in H. Keller and G. Ulfstein (eds), *UN Human Rights Treaty Bodies: Law and Legitimacy* (2012) 356, at 374ff.

[18] See only Murdie and David, 'Shaming and Blaming: Using Events Data to Assess the Impact of Human Rights INGOs', 56 *ISQ* (2012) 1; Simmons, *supra* note 3; T. Risse, S. Ropp and K. Sikkink (eds), *The Power of Human Rights: International Norms and Domestic Change* (1999).

[19] Shikhelman, *supra* note 1, at 764. It is unclear, though, what the related variable precisely measures: the text says 'number of NGOs in a state per capita' (*ibid.*) (all NGOs or only human rights NGOs?), while the purported data source in note 79 employs a 'count measure of country memberships in I[nternational] N[on-]G[overnmental]O[rganization]s in a given year (i.e., the number of INGOs citizens of a state have membership in'. Hafner, Burton and Tsutsui, 'Human Rights in a Globalizing World: The Paradox of Empty Promises', 110 *American Journal of Sociology* (2005) 1373, at 1393. These are not identical data and they may implicate different causal mechanisms (such as non-governmental organization (NGO) pressure from within versus NGO pressure from without).

system accords some consequential role to treaty body views in judicial proceedings (which is rarely the case).[20] Some enforcement mechanisms, such as political lobbying and naming and shaming, are available irrespective of legal status, but others may be either limited to legally binding obligations or may work differently depending on their legal status. It would have enriched the argument and analysis if these differences had been explored and clarified upfront.

Importantly, at the normative level, legal non-bindingness does not need to equate with non-bindingness as such. Indeed, we disagree with those that conceptualize bindingness in a strictly dichotomous fashion solely in terms of legal (non-)bindingness[21] and, instead, believe that it is both theoretically correct and empirically more fruitful to view bindingness as occurring in qualitatively different degrees, with legal bindingness being only one manifestation of bindingness. Viewed this way, the absence of legal bindingness does not imply non-bindingness, only legal non-bindingness, apart from which there are arguably other forms of non-legal bindingness that can be political, social or moral in nature. In the last instance, the existence of the bindingness of a norm or decision depends on the intersubjective recognition that such a norm or decision shall act as a constraint on one's choices: while the strength of such recognition and of the consequent constraining effect may differ, few would assert that commitments and obligations that are not legally binding, therefore, *ipso facto*, leave the issues to which they relate fully and unconditionally within the discretion of those involved; rather, they usually create expectations of adherence and compliance, constrain decision-making and, to this extent, exert binding effect, even if not legally enforceable.[22] Shikhelman also seems to espouse such a graduated understanding of different degrees of bindingness when she describes soft law as 'quasi-legal norms that do not have a *completely* [but thus apparently *some*] binding force'.[23] This 'incompletely' binding force and its expected effects on outcomes would have merited additional theoretical examination.

We may safely expect that the degree of explicitly or implicitly recognized bindingness of adverse HRC views will vary, *inter alia*, with regime type. For liberal democracies, the legally non-binding HRC views will tend to have greater constraining effects than they do for autocracies, and, to this extent, they will be more binding, raising the justificatory hurdles for any non-compliance. Hypothesizing variable degrees of bindingness thus lends support to Shikhelman's first hypothesis that democracies should be more likely to comply with HRC views than autocracies.[24] As for the auxiliary expectation that compliance by autocracies might be driven in particular by efforts

[20] Kanetake, 'UN Human Rights Treaty Monitoring Bodies before Domestic Courts', 67 *International and Comparative Law Quarterly* (2018) 201, at 217–218; van Alebeek and Nollkaemper, *supra* note 17, at 362–367.

[21] See Broude and Shereshevsky, *supra* note 13.

[22] As Oscar Schachter had long ago aptly noted in the context of non-binding agreements, the fact that 'noncompliance by a party would not be a ground for a claim for reparation or for judicial remedies … is quite different from stating that the agreement need not be observed or that the parties are free to act as if there were no such agreement'. Schachter, 'The Twilight Existence of Nonbinding International Agreements', 71 *AJIL* (1977) 296, at 300.

[23] Shikhelman, *supra* note 1, at 754 (emphasis added).

[24] *Ibid.*

to increase their international reputation, we would add that reputational motives matter for democracies as well as that they should likewise be anxious not to damage their reputation by failing to comply with the output of treaty body procedures that they have freely and voluntarily accepted.[25]

3 Data Issues

Shikhelman's approach of using the grades from the follow-up reports to establish compliance with the HRC's views is intriguing. It allows the researcher to assess compliance on a case-by-case basis without the strenuous task of having to gather information on the implementation of every case all by oneself. However, reliance on those 'scorecards' comes with a number of difficulties that need to be dealt with. We can distinguish between two types of possible selection effects stemming from (i) the possibly inconsistent behaviour of an understaffed committee and the Office of the UN High Commissioner for Human Rights' Secretariat and (ii) differing reporting behaviour by respondent states. In what follows, we discuss the ramifications of those possible sources of bias in more detail.

Although Shikhelman mentions early in her article that 'states do not always receive letters reminding them to report the status of implementations', this fact is later all but ignored.[26] This is problematic since the selection of cases that are followed up by the committee might not be random. For example, within Shikhelman's observation period from 2014 to 2016, the HRC discussed 90 cases with views dating back as far as 1989. From the first view in 1977 until 2013, the year before the observation period starts, the HRC had issued 834 adverse views.[27] Even if the committee had reviewed the implementation status of 30 additional views every year since the position of a follow-up rapporteur was first created in 1990,[28] it would have covered only 690 of them until 2013, leaving 144 adverse views without formal follow-up and compliance information.

Although the committee notes that 'follow-up information has been systematically requested in respect of all views with a finding of a violation of the Covenant',[29] it often cannot keep up with assessing implementation after the initial request. Unfortunately, we simply do not know on what grounds a case is being picked, or not being picked,

[25] One might object that non-compliance with adverse views will not tarnish most democracies' otherwise solid human rights reputation, yet if that were correct then one should also not expect that occasional compliance with adverse views by autocracies will have much of an influence on their reputation if they 'continu[e] to violate human rights in general' (*ibid.*). The relevant audiences and stakeholders should be able to see through such token attempts to embellish a state's otherwise poor human rights record.

[26] Shikhelman, *supra* note 1, at 762.

[27] These numbers are taken from a dataset we are currently compiling in order to assess compliance with all views/decisions by all United Nations treaty bodies with an active individual complaints procedure.

[28] Schmidt, 'Follow-Up Mechanisms before UN Human Rights Treaty Bodies and the UN Mechanisms Beyond', in A. Bayefsky (ed.), *The UN Human Rights Treaty System in the 21st Century* (2000) 233, at 235.

[29] Human Rights Committee (HRC), *Annual Report* (1994), Doc. A/49/40, vol. 1, at 84.

to be followed up in the reports. An illustration of the arbitrariness of this process is provided by the fact that some dialogues remain ongoing for decades (without considerable new evidence by the parties),[30] while others are closed with an unsatisfactory finding after a short while.[31] As a result, the distribution of cases across states in Shikhelman's sample is not representative of the distribution of views among all states subject to the Optional Protocol to the International Covenant on Civil and Political Rights (ICCPR).[32] Figure 1 illustrates this, showing the 30 countries that received the most views since 1977 and the number of views of those countries in Shikhelman's sample. Although Bosnia and Herzegovina is the country with the highest number of views reviewed as part of the HRC's follow-up procedure during the sample period, it only ranks 21st among the countries that received the most views overall. At the top of the list are South Korea, Jamaica and Belarus, none of which make an appearance in Shikhelman's dataset.

One cannot help but get the impression that the processing of views in the follow-up procedure does not occur systematically but, rather, selectively and based on the availability of incoming information. In light of the secretary's underfunding, this is understandable, but with respect to the randomization, and thus representativeness,

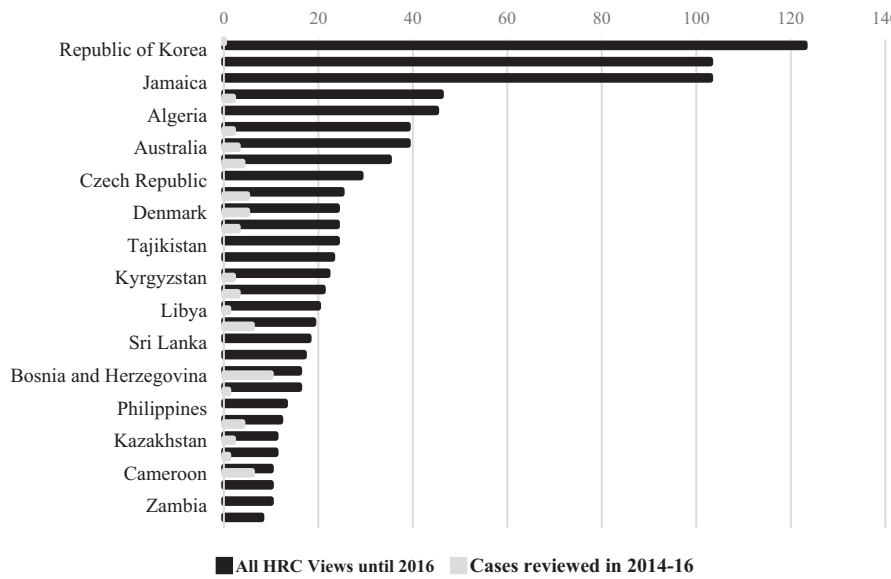

Figure 1: *Distribution of HRC views: Overall versus Shikhelman's sample*
Note: This chart only reports the first 30 countries with the most views overall and does not provide an exhaustive list of all HRC views rendered before 2016 or include all of the views in Shikhelman's sample.

[30] See, e.g., HRC, Views on Communication no. 45/1979 (*Suarez de Guerrero v. Colombia*), which have remained under follow-up review since the latter began, with the last assessment having been made in 2013.

[31] See, e.g., HRC, Views on Communications no. 1108/2002 and 1121/2002 (*Karimov and Nursatov v. Tajikistan*).

[32] Optional Protocol to the International Covenant on Civil and Political Rights 1966, 999 UNTS 171.

of the sample, this is problematic as it suggests a source of serious bias. The fact that the follow-up reports normally concentrate on multiple communications concerning only a few countries and keep reviewing the same communications from report to report might suggest that Shikhelman only ended up with these 28 countries because they were the ones on top of the paper pile on the desk of the committee member responsible for follow-ups. In other words, we cannot say what factors influenced the timing of the committee's follow-up review.[33] If one relies only on the follow-up assessments without controlling for such possible factors, however, one is likely to end up with a non-randomly drawn sample. Moreover, grading takes place based on the state's reply, but the HRC often keeps the dialogue open to wait for the author's reply or even to meet again with representatives of the state party. Information obtained through these replies or meetings may lead to changes in the initial grades.[34] This implies that, as long as the dialogue is not closed, we cannot necessarily take any individual assessment as being final. Within Shikhelman's observation period, 76 cases received grades, but only 10 of them were closed with a finding of non-, partial or full compliance. To make things even more complicated, the HRC sometimes closes dialogues without giving grades at all.[35] It would be worthwhile to know whether such communications have ended up in Shikhelman's sample or not.

While Shikhelman does not address selection effects that might emerge on the part of the committee, she rightly draws the reader's attention to possible systematic differences between countries that submit follow-up reports and those that do not. In other words, if all the countries in her sample only end up there because they all have common traits that set them apart from other countries, then the sample would likewise not be random. Shikhelman posits that a country's degree of democracy and its human rights performance are characteristics that may influence that country's propensity to send a reply letter, thus ending up in her sample. To check for such bias, she conducts *t*-tests for the difference in means between her current dataset and a dataset that covers the views issued over the years 1997 to 2013 with respect to respondent countries' Polity IV regime-type scores and Christopher Fariss' human rights scores as the dependent variables. The rationale behind this is simple: if there is no statistically significant difference between her sample (containing only responding states) and the old dataset (which also includes non-responding states) concerning these two variables, then there is no 'self-selection' of states into the sample.

[33] Indeed, even a top-ranking member of the HRC in a recent conversation with us could not pin down the factors that determine a communication's likelihood to be reviewed.

[34] See, e.g., HRC, Views on Communications no. 2094/2011 (*F.K.A.G. et al. v. Australia*); HRC, Views on Communications no. 2136/2012 (*M.M.M. et al. v. Australia*), which initially received 'C2' grades for non-repetition (see HRC, *Follow-Up Report* (2016), Doc. CCPR/C/118/3, 2), a grade that was later changed to 'E' after the author's counsel responded to the state reply (HRC, *Follow-Up Report* (2017), Doc. CCPR/C/119/3, 9). There are numerous other examples for this, e.g., regarding HRC, Views on Communications no. 1917/2009 1918/2009, 1925/2009 and 1953/2010 (*Prutina et al. v. Bosnia and Herzegovina*), no. 2091/2011 (*A.H.G. v. Canada*), no. 2008/2010 (*Ali Aarrass v. Spain*), no. 2077/2011 (*A[ng] S[herpa] v. Nepal*).

[35] See, e.g., HRC, Views on Communication no. 1153/2003 (*Llantoy Huamán v. Peru*); HRC, *Follow-Up Report* (2016), Doc. CCPR/C/118/3, 33.

However, there are some problems with this rationale. First, other factors come to mind that might also determine a country's response behaviour, such as its resources that are available for producing innumerable state replies to a growing number of treaty body requests.[36] It is not clear why selection bias should primarily occur along the lines of regime type and human rights performance. Second, we take from Shikhelman's wording that data in both datasets is at the level of communications, not countries. If that is the case, then data for both indicators are likely skewed due to the uneven distribution of communications among countries, which in turn violates the normality assumption underlying the *t*-test.[37] Third, the 'old dataset' that she uses likely includes, among others, the same countries that ended up in her sample as well as those that generally reply but which did not receive a view within her observation period. Hence, the two groups she is comparing seem to be unrelated only as long as one takes the communications as the unit of analysis.

Both the Polity IV regime-type score and the human rights score are measured at the country level; however, the units of observation within the two groups are communications, and the latter are not independent from each other. In the end, Shikhelman does not assess the 'difference between reporting states with non-reporting states' but, rather, between a set of states at one point in time and another – partially overlapping – set of states at an earlier point in time.[38] Even assuming that this approach suffices to establish randomness, the fact that she does not tell the reader at what level she measures her variables reduces the plausibility and external validity of her approach. A better way to control for bias through self-selection would be to run a Heckman selection model, which consists of a selection equation that considers variables that predict the propensity of each observation to get selected in order to then incorporate this information in the main regression.[39]

Given these caveats, Shikhelman's assumption that one can obtain a non-biased, random sample by using follow-up reports stands on very shaky ground, with all of the negative implications for statistical inference that non-randomness implies.

4 Methodological Issues

We have argued in the previous part that relying solely on the grades included in the HRC's follow-up reports likely implies considerable selection bias. In addition to these

[36] Creamer and Simmons, in 'Ratification, Reporting, and Rights: Quality of Participation in the Convention against Torture', 37 *Human Rights Quarterly* (2015) 579, find that institutional capacity heavily conditions country report submission under the Committee against Torture. Cole, 'Mind the Gap: State Capacity and the Implementation of Human Rights Treaties', 69 *IO* (2015) 40, further finds that state capacity, such as bureaucratic efficiency, plays an important role in a state's ability to implement views by the HRC; see also Grewal and Voeten, 'Are New Democracies Better Human Rights Compliers?', 69 *IO* (2015) 497, for similar arguments in the context of compliance with the judgments of the European Court of Human Rights.

[37] A Shapiro-Wilk-Normality Test of the human rights score in Shikhelman's 'new dataset' – the only dataset available to us – yielded a p-value of 0.00035, indicating high deviance from a normal distribution; there is no indication in the article that the problem of skewness has been addressed, for example, through logarithmic transformation.

[38] Shikhelman, *supra* note 1, at 767.

[39] See, e.g., Cole, *supra* note 8, for such use of a Heckman selection model in a related research context.

data issues, we have a few methodological concerns. To begin with, it is the categorization of different remedies employed in the reports that merits particular scrutiny. Shikhelman uncritically adopts the classification found in the reports for the operationalization of both her dependent variable as well as one independent variable, and thus unnecessarily reduces the content validity of her measurements. Why should the categories that the HRC uses for its remedies be problematic? Simply because they were not defined with a view to being used as values of a nominally scaled variable. For example, we do not know the rationale behind the category 'effective remedy', but, if one looks it up in the reports, it becomes clear that this category most of the time entails several different remedies, such as 'monetary compensation' or 'release from imprisonment'. However, if they are lumped together in the 'effective remedy' category, these different remedies will only be graded as one. In short, 'effective remedy' is not a clear-cut category and that is a problem because, would it have been phrased as distinct remedies, there would be a different outcome. For the purposes of illustration, take Communication no. 1107/2002 (*El Ghar v. Libya*) from Shikhelman's sample in which the follow-up report finds the respondent state to be compliant with some of the remedies mentioned under the 'effective remedy' umbrella and gives it the letter grade 'B'. If Shikhelman had disentangled the 'effective remedy' category, she would instead have ended up with an 'A' rating for one remedy (issue of passport) and a 'C' rating for the other remedy (compensation). Another peculiar category is 'non-repetition', which typically involves general measures such as legislative changes that may also exist as separate remedies. The problem here is that the HRC created categories that are not mutually exclusive, with the consequence that they sometimes substantively overlap and thus do not represent distinct, separate values of the 'remedy' variable.

To illustrate the resulting problems, we selected all communications for which the remedial measured had been graded in the follow-up reports of Sessions 112–118 (this covers the years 2014–2016) from a dataset including all HRC views issued between 1979 and 2019.[40] We then employed a typology for remedies that takes its cues from the supervisory practice of the Council of Europe's Committee of Ministers with respect to the execution of the judgments of the European Court of Human Rights,[41] which consists of three types: (financial) compensation, individual measures and general measures (see Table 1).

This typology is based on our assumptions concerning the different remedies' likelihood of implementation. Comparatively, the payment of (relatively small amounts of) monetary compensation should generally be the least costly remedy for states,

[40] It should be noted that our sample likely differs from Shikhelman's because it does not include those views reviewed again at later HRC sessions beyond Shikhelman's period of observation (which covers Sessions 110–118), while it does include one observation – HRC, Views on Communication no. 1153/2003 (*Karen Noelia Llantoy Huamán v. Peru*) – that was closed with a finding of satisfactory implementation, but without an expressly stated remedy-specific grade. Since the complainant confirmed that the respondent state had fully complied with the required compensation remedy, we assigned a grade of 'A' to that remedy.

[41] See Rule 6, paragraph 2, of the Rules of the Committee of Ministers for the Supervision of the Execution of Judgments and of the Terms of Friendly Settlements, available at https://rm.coe.int/16806eebf0.

Table 1: Different types of remedies

Category	Compensation	Individual measures	General measures
Examples	'appropriate compensation'	'release under individually appropriate conditions for those authors still in detention' 'rehabilitation' 'continuing its efforts to establish the fate or whereabouts of the victim' 'bringing to justice those responsible' 'necessary psychological rehabilitation and medical care'	'non-repetition' 'abolishing the obligation for family members to declare their missing relatives dead to benefit from social allowances'

especially if it is the only measure required, enabling a state to end a case without generating much noise in the public sphere. The fact that Shikhelman finds the compliance pattern for compensation remedies to be different from other individual remedies seems to support our assumptions, although her findings counter-intuitively suggest a low compliance rate with this sort of remedy. We largely agree with Shikhelman concerning her conjecture about the compliance costs of individual versus general remedies, but we trace compliance with distinct general measures separately. This is necessary since the committee's non-repetition admonishment 'to take steps to prevent similar violations in the future' is so generally phrased that it is unclear what measures the respondent state has to take and how costly these might be or, indeed, if any meaningful general measures have to be taken at all beyond the (comparatively costless) application of existing rules and regulations in a covenant-compliant manner.[42] In short, assessing and differentiating the costliness of individual, compared to general, measures is empirically not as clear-cut as the conceptual distinction may suggest. In addition, we do not account for the 'publication of views' remedy because it seems to involve few compliance hurdles and, thus, is not particularly informative. Figure 2 shows the frequency of follow-up grades and their distribution across the three categories.

Since one category can contain several grades for different remedies, we coded all of the remedies within that category and then calculated the mean for that category in such a case. By this, we constructed a continuous compliance variable with 2 as the highest value (indicating full compliance, or perfect 'A' grades) and –2 as the lowest value (indicating measures in contrary to the recommendations of the HRC or perfect 'E' grades). The overall frequency of the grades seems to coincide with Shikhelman's assessments, with 'C' being the most frequent and 'D' being the least frequent grade. However, breaking down remedies into three categories allows us to compare the compliance patterns between them. Figure 2 shows that – contrary to our expectations but in line with Shikhelman's findings – compensation remedies are the ones that are least complied with (mean = 0.22). This might be because the views do not

[42] HRC, Views on Communication no. 2149/2012 (*M.I. v. Sweden*), para. 9.

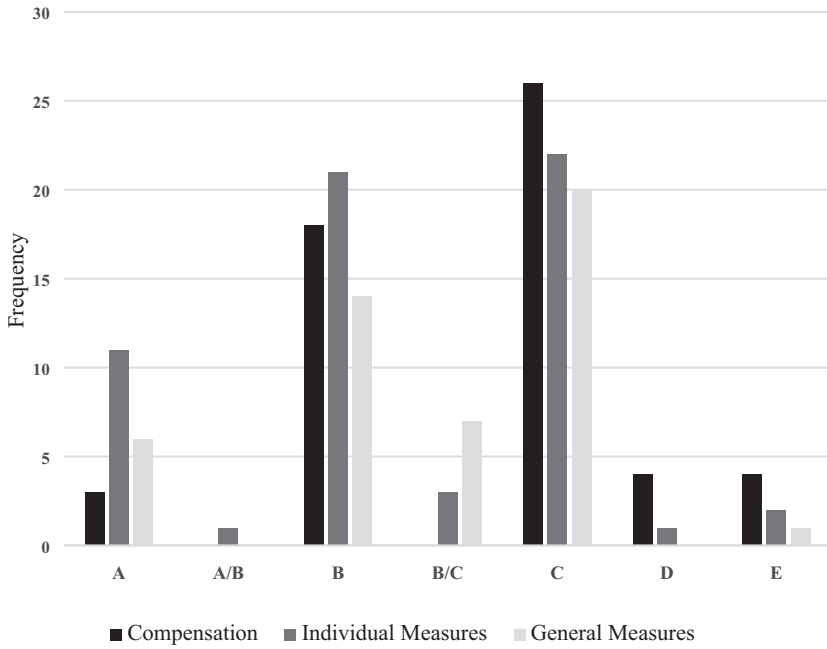

Figure 2: *Frequency of follow-up grades by type of remedy*

specify a specific sum to be paid to the victim, which might confuse the respondent state as to what is required of it. The bar plot further reveals that individual measures are the ones most likely to be implemented (mean = 0.67), closely followed by general remedies (mean = 0.57). Although this is in line with our expectations, the high compliance rate with general measures is surprising. Similarly striking is that general measures are nearly always addressed in some way and virtually never ignored or reversed (grades 'D' or 'E').

We would simply like to highlight here that assigning the many different specific remedies to a parsimonious, theoretically grounded and clear-cut set of types allows for a more targeted analysis of the effects of various factors on compliance with these remedies. Interestingly enough, our three-tiered typology is largely in accordance with Shikhelman's own hypothesis concerning compliance with different types of remedies. Later on in her article, however, it is not made clear how these general theoretical expectations are operationalized in the context of her empirical analysis. Figure 2 in her article provides a range of sometimes more, sometimes less, specific remedies (compare, for example, 'release from imprisonment' with 'effective remedy' or 'other'), mixing in the process individual and general measures, without telling the reader which ones are which. The careful allocation of stipulated remedies to different types, and possibly even differentiating them further beyond individual and general measures,[43] matters not least since we can expect that the implementation of different remedies may well be subject to different causal mechanisms.

[43] It might indeed be more insightful to conceptualize compliance with the different remedies as dependent variables, rather than using a binary distinction between individual and general measures as an independent variable.

The same problems arise in the categories that Shikhelman finds for different subject matters. While the argument that a communication's subject matter is relevant for the likelihood of its implementation is in principle plausible, the categories do not result directly from the communications. For example, is a view that concerns deportation to another country where a person may be in danger of being tortured because he or she is a gay member of an ethnic minority an 'immigration', a 'lesbian, gay, bisexual or transgender' or a 'minority rights' case? An alternative would be to focus on the specific ICCPR articles that are found to have been violated in the views in the dataset. From the Committee against Torture, for example, we know that many cases concern the non-refoulement provision of Article 3 of the UN Convention against Torture (CAT), and while these would likely be classified as immigration cases in Shikhelman's scheme with a low expected probability of implementation, the compliance rate with those CAT decisions is surprisingly high, at around 70 per cent.[44] There is no *a priori* reason to anticipate, *ceteris paribus*, that similar HRC findings under Article 7 of the ICCPR, the covenant's anti-torture provision, should yield systematically different results.

In addition to the difficulty of properly categorizing the different remedies, Shikhelman faces the challenge of handling the hierarchical structure of her data. Although she never explicitly problematizes it, Shikhelman's variables are measured on three different levels: the country level, the communication level and the remedy level. This special data structure needs to be addressed in the regression by way of an adequate multilevel model. However, Shikhelman is essentially fitting a two-level model to three-level data since she only accounts for correlations between communications and not additionally between countries. This error will lead her to misattribute observed variation only to the two included levels and underestimate the effects of country attributes. What is more, including independent variables on the country level in the remedies-within-communications model, whilst ignoring 'country' as a level in the model, may lead Shikhelman to underestimate the standard errors of the country-level variables and, as a result, put her at risk of making Type 1 errors.[45] Such problems can be dealt with by generating a mixed-effects model that incorporates random effects on all three levels. Apart from that, it is worth mentioning that Shikhelman measures most of her independent variables at the country level. Thus, a good proportion of her interest is in explaining variation across countries and not between communications or remedies. However, in her sample, there are only 28 sources of variation for these variables, leaving her with few degrees of freedom. Indeed, had she not included the 'individual remedy' variable, Shikhelman would basically only compare 76 cases. Hence, variation across the 300 observations is very limited, which in turn minimizes the generalizability of her results.

Probably the most important limitation in Shikhelman's approach, however, is her choice of cross-sectional analysis over event history analysis, but when it comes to

[44] Von Staden, '(Sometimes) as Good as a Court: Compliance with the Individual Complaints Decisions of the UN Committee against Torture', working paper (on file with authors). Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment 1984, 1465 UNTS 85.

[45] J. Hox, M. Moerbeek and R. de van Schoot, *Multilevel Analysis*: *Techniques and Applications* (2018).

compliance with adverse views, time is clearly of the essence. By not taking time into account when assessing compliance, one inevitably makes the assumption that all implemented remedies are qualitatively the same, regardless of the time that has passed since the finding of the violation. To illustrate, take the hypothetical scenario in which a political opponent of a government remains imprisoned for 20 years after the HRC has found the incarceration to violate Articles 10 and 26 of the ICCPR. And it is only after, say, a change of government that the respondent state finally implements the remedies called for in the initial view and frees that person, but without letting the HRC know of its actions. Now imagine the same case, only this time the government releases the political prisoner shortly after it received the HRC's view – for whatever reasons. In the end, both cases will get the letter grade A in the follow-up report, but are they really qualitatively the same? The perception that justice delayed is justice denied has led some scholars interested in assessing compliance with the judgments of the European Court of Human Rights to take the time between the Court's judgment and the implementation as their key dependent variable.[46] There is no compelling theoretical or methodological reason why the time dimension should not also be included when studying compliance with the views of the HRC.

Also, note that in the first scenario outlined above the likelihood of compliance changed considerably after regime change. With Shikhelman's cross-sectional approach, however, this crucial information is lost because we only have information on regime type at one point in time. In a case like this, however, within-country variance across time is considerably more informative as to the factors that cause compliance than variance between countries. The best way to incorporate the effects of time in a statistical analysis is the use of panel data. It is now a common approach in the social sciences to use statistical models called 'event history' or 'survival' models in order to make sense of right-censored panel data when the dependent variable is a sort of event or categorical variable.[47] The problem of right censoring occurs when an observed unit does not experience the event of interest – which, here, is an assessment indicating compliance in the HRC's follow-up reports – within the observation period. Cross-sectional regression models wrongly treat such observations as conclusively having failed to experience the event of interest by the end of the observation period. This is why many researchers who do not employ survival models simply eliminate such observations, with the undesirable consequence of losing all of the information contained in them. In Shikhelman's case, there are numerous communications that are mentioned in the follow-up reports, but they do not receive remedy-level grades and, hence, do not end up in her sample.[48] However, they might receive a decision shortly after the observation

[46] See Grewal and Voeten, *supra* note 36; Anagnostou and Mungiu-Pippidi, 'Domestic Implementation of Human Rights Judgments in Europe: Legal Infrastructure and Government Effectiveness Matter', 25 *EJIL* (2014) 205.

[47] See, e.g., P. Allison, *Event History and Survival Analysis* (2014); Box-Steffensmeier and Jones, 'Time Is of the Essence: Event History Models in Political Science', 41 *American Journal of Political Science* (1997) 1414.

[48] For example, the last follow-up report considered by Shikhelman, adopted at the HRC's 118th session – UN Doc. CCPR/C/118/3, 1 August 2016 – contains numerous instances of this sort, such as the entries relating to HRC, Views on Communications no. 1353/2005 (*Afuson Njaru v. Cameroon*), no. 1620/2007 (*J.O. v. France*), no. 1376/2005 (*Bandaranayake v. Sri Lanka*), no. 2179/2012 (*Young-kwan Kim et al. v. Republic of Korea*), no. 2051/2011 (*Basnet v. Nepal*) and several others.

period. Simply disregarding these observations results in losing the information that they can provide. Survival models make use of information included in these right-censored observations while also accounting for the fact that the likelihood of a satisfactory decision changes with the time passed since the publication of the view.

In order to illustrate what can be achieved by using event data, we added a dichotomous compliance variable as well as some of the independent variables used by Shikhelman to our sample of HRC views followed up and graded between 2014 and 2016.[49] Figure 3 depicts the probability of non-compliance with adverse views – that is, of 'surviving' without compliance – over time, with 'time' measuring the number of years from the year in which a view was originally issued.[50] The graph on the left shows that democratic states are not only more likely to comply generally, but that they also adopt remedial measures more quickly than non-democracies in that their mean probability of survival is consistently lower than that of non-democracies and approaches zero about 10 years after the views in the data set had been issued.[51] In other words, by that time, democracies will have implemented at least some sufficiently compliant remedial measures, whereas non-democracies still persist thereafter with no such measures having been taken. The graph on the right is equally indicative of the importance of time for compliance.[52] It shows that the probability of (continued) non-compliance initially drops noticeably faster for individual measures than for general and compensation

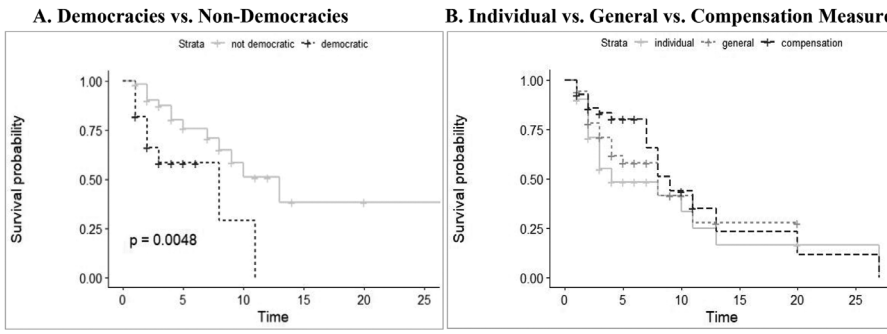

Figure 3: *Kaplan-Meier survival estimates of human rights committee views on individual communications and remedies, by regime type and type of remedy*

[49] In Figure 3.A, compliance is measured at the level of adverse views. We coded the compliance status of a view as '1' when 50 per cent or more of the required remedies received at least a 'B' grade. Although this admittedly measures not only full but also partial compliance, it here solely serves the illustrative purpose of indicating that a respondent state has at least taken some of the required remedial measures (considering only full compliance and the implementation of all remedies would have simply provided too few data points because of the small sample size). It should also be noted that our sample is likely to be biased because of the selection effects discussed earlier.

[50] A communication enters the risk set in the same year that the adverse view was issued.

[51] To assess democracy, we coded countries with a V-Dem liberal democracy score higher than 0.7 as 'democratic' and countries with lower scores as 'non-democratic'. See Coppedge *et al.*, V-Dem Country-Year Dataset, vol. 10, available at https://www.v-dem.net/en/data/data-version-10/.

[52] In Figure 3.B, compliance was coded at the level of remedies, with '1' indicating an 'A' or 'B' grading and '0' indicating a 'C', 'D' or 'E' grading for a required remedy.

measures, so that five years after an adverse view has been issued, individual measures have a 55 per cent risk of being complied with, while general remedies have approximately a 35 per cent risk of being implemented, and compensation remedies face only a 20 per cent risk of implementation. However, after 10 years, all three types of remedies have become subject to essentially the same hazard of being implemented (around 60 per cent). Somewhat surprisingly, 20 years after the original view, it is compensation remedies that show the highest compliance hazard among the three types.

While this indeed suggests that individual remedies are more readily implemented initially, as both Shikhelman and we would expect from a theoretical point of view, general measures catch up thereafter.[53] One intuitive explanation for this pattern is the circumstance that most general remedies simply take longer to implement. Notwithstanding their limited generalizability due to the small sample size, these findings stress the importance of properly incorporating time as a mediating factor in the analysis of compliance with treaty body views. Certainly, using survival analysis as the method of choice does not come without its own challenges as it requires a sufficiently long observation period to avoid left censoring and to create enough variance over time. This will inevitably pose new problems due to the changing practices of the HRC in handling the follow-up procedure. It would lead, however, to insights that are more generalizable by taking into account the full range of political processes behind compliance.

5 Conclusion

Vera Shikhelman's article is a welcome contribution to research exploring why states do (or do not) comply with adverse HRC views. But, as we have sought to indicate in this contribution, the devil is in the detail, requiring a number of consequential theoretical and methodological choices that have implications for the internal and external validity of such research. We believe that the choices made in 'Implementing Decisions of International Human Rights Institutions' concerning some of the theoretical expectations, the handling and use of the available data and the methodological set-up and execution of the study do not yet maximize the causal inferences that one could draw from the available data and the information contained therein. However, as an old Chinese proverb reminds us, every journey of a thousand miles begins with a single step. Vera Shikhelman has taken that first step, and we look forward to further interdisciplinary research on the important topic of compliance with treaty body decisions in individual communication procedures, procedures that contain for many complainants the sole hope of improving their human rights situation and, for that reason alone, deserve to be understood better.

***

Vera Shikhelman continues the debate with a Rejoinder on our EJIL: Talk! blog.

[53] If we had more data, we might even observe a higher risk of implementation for general remedies after 20 years – which, in turn, could explain the negative coefficients for the 'individual remedy' variable in Shikhelman's regression results.