Journal of Public Policy (2018), 38:3, 275–304 © Cambridge University Press, 2017
doi:10.1017/S0143814X16000271

Assessing the validity and reliability of measurements when evaluating public policy

MICHELE CREPAZ
Department of Political Science, Trinity College Dublin, Ireland
E-mail: crepazm@tcd.ie

RAJ CHARI
Department of Political Science, Trinity College Dublin, Ireland
E-mail: charir@tcd.ie

Abstract: A substantial aspect of scientific research involves linking concepts to observations using measurements. This exercise has raised questions among researchers of whether or not measurements "truly" and "reliably" capture ideas and observations. We address this question by setting out a methodological standard on how to assess the validity and reliability of measurements. We do this by examining measurements that evaluate public policy, arguing that this topic is gaining increasing attention from political science researchers and policymakers. The analysis concerns measurements of the level of transparency and accountability of lobbying laws, central to recent regulatory policy research. We conduct convergent validation, content validation and reproducibility tests on four indices applied to 13 regulations found worldwide. By doing so, the article provides scholars with an evaluation of measurements of lobbying laws' robustness, while offering methodological and theoretical lessons of value to larger regulatory and public policy scholarship.

Key words: lobbying, measurement, regulation, reliability, validity

Introduction

A substantial aspect of the scientific method of investigation involves *concept formation*, the *development of operationalisation* and *scoring*. Adcock and Collier define a *systemised concept* as a "specific formulation of a concept used by a given scholar or groups of scholars involving a specific definition"; *operationalisation* includes "developing, on the basis of a systemized concept, one or more indicators for scoring/classifying cases (where indicators are also referred to as ‘measures’)”; and *scoring* cases involves “the application of indicators to produce scores” (2001, 531).[1]

Researchers naturally question whether existing operationalisations “truly” (or meaningfully) reflect the concept they seek to measure (King et al. 1994). They also question whether or not elements of arbitrariness accompany the act of measuring.

Adcock and Collier (2001) addressed these concerns with respect to measurement validity. They offer a methodological standard to assess measurement validity departing from the discussion of the formation of systematised concepts to the analysis of several methods of validation (content, convergent and construct validation). Krippendorff (2004) dedicated attention to the issue of arbitrariness in scoring procedures with implications on the reliability of measurements.

Since these groundbreaking studies, researchers have paid increasing attention to the evaluation of concept formation, measurement validity and measurement reliability in political science: Maggetti and Gilardi (2016) studied measurements of policy diffusion; McMenamin (2004) analysed measurements of the varieties of capitalism; Rogers and Weller (2014) investigated indices of state capacity; Rocco and Thurston (2014) focussed on measures of institutional change; and Munck and Verkuilen (2002) as well as Seawright and Collier (2014) evaluated different indices of the level of democracy.

Policy scholars, in contrast, have conducted surprisingly little research on different measurements of public policy outputs. Notable exceptions are seen in the work of Rosenson (2003) who developed a measurement of the stringency of ethics policy, Witko (2007) who produced an index of the rigour of campaign finance regulations and Wagenaar et al. (2005) who focussed on measures of the strictness of alcohol restriction policy. In these studies, however, the scholars fail to engage with the larger methodological debate about measurement validity and reliability. With this article, we hope to help fill this gap in the literature of policy analysis by proposing methodological standards about measurement validity and reliability that can be applied to public policy research. This is important because our ideas can be applied to any measurement that evaluates public policy based on a coding procedure, and thus add rigour when evaluating policy from a more empirical perspective.

To this end, we decide to focus on measurements of the levels of transparency and accountability of lobbying laws. This regulatory policy falls within the family of ethics, integrity and transparency laws alongside initiatives such as Freedom of Information (FoI) laws (Banisar 2006). However, although FoI requests allow citizens to better understand why

[1] See also Munck and Verkuilen (2002, 8) for a framework of the analysis of data, which highlights the same points, although with slightly different terms.

decisions have been made by the state, effectively regulating the actions of public actors and holding them accountable for their actions, lobbying laws regulate the relationship between private actors who are seeking to influence the state. Such laws help to prevent corruption and shed light over the participation of interest groups in the policymaking process. The increasing significance of such laws is reflected in the fact that the number of countries worldwide that have established lobbying laws in the 2000s is almost three times that had rules in the 1900s.

A fundamental dimension of lobbying rules is that lobbyists must register with the state, usually an independent regulator, before contact is made with elected officials and high-level civil servants that are targeted. The legislation defines which lobbyists are regulated (such as consultancies, in-house corporate lobbyists, nongovernmental organisations (NGOs) and professional associations) and which organisations may be exempt from registration (such as charities). The amount of information that lobbyists must disclose varies among jurisdictions with lobbying laws. This ranges from simply stating the name of the bill, ministry and official being targeted to disclosing more detailed information on the money spent on lobbying. The regulator generally publishes the registration information in an open, online database, allowing citizens to access it and see who is lobbying whom and about what. If there is a breach of rules – say that a lobbyist is found to be active without being registered – the regulator may impose sanctions such as fines or imprisonment. In order to prevent potential conflicts of interest, many lobbying laws also have "revolving door" (or "cooling-off") provisions, which stipulate the time period that public officials leaving office have to wait before entering the lobbying industry.

The provisions in lobbying laws regarding these conceptual dimensions can be more or less *robust*. For example, disclosure requirements for lobbyists or sanctions in cases of misconduct can be more or less strict. *Robustness* is thus identified as the strictness of the regulation, or, more precisely, the level of transparency and accountability that lobbying laws can guarantee.

We argue that the concept of the *robustness of lobbying laws* is an (increasingly) "essentially contested concept" (Gallie 1955). To be clear, we do not seek to claim that the concept of robust lobbying laws is an equally "essentially contested concept" such as democracy, for example, debated in the literature for decades (Munck and Verkuilen 2002; Bowman et al. 2005). Nor do we seek to enter into the debate that belongs better in political theory on how to best conceptualise an "essentially contested concept" itself (Connolly 1974; Clarke 1979).

We simply seek to argue that, from an objective point of view, the debate on the conceptualisation of robust lobbying regulations that started over

25 years ago shows evidence of fulfilling Gaillie's original criteria in order to be considered "contested". The first criterion is *appraisiveness*, ascribing to the concept "some kind of value achievement" (Gallie 1955, 71). The last 20 years of regulatory policy research has witnessed a booming popularity of a plethora of studies analysing lobbying rules and evaluating their robustness as the number of countries adopting laws has increased.[2] In these studies, the positive normative valance attached to the concept of robustness stimulates an ongoing debate around the adoption of lobbying laws in contemporary democracies.

The second criteria, *internal complexity*, would signify that different users might view or describe robustness differently, consisting of different dimensions (Collier et al. 2006; Chambers and Carver 2008). As seen later, when compared with the pioneers originally writing on the theme, the more recent authors have described robustness differently and attributed more dimensions to the construction of their indices measuring the legislation's robustness.

Third, the concept is *open* (Gallie 1955; Chambers and Carver 2008, 257), where its meaning is periodically under review and constant interpretation. The literature on lobbying has produced six measurements of the concept of *robustness* of lobbying laws, four of which are studied in this article.[3] The first is that developed by Opheim (1991) at the United States (US) state level, followed by Newmark (2005) who revised Opheim's conceptual dimensions. Chari et al. (2010) extended the concept to the global comparative analysis of lobbying laws. Moreover, two years later, Holman and Luneburg (2012) produced modifications to the conceptual dimensions of robustness, where subsequent scholars more explicitly developed an empirical index on the basis of their work (Crepaz 2016a). These pieces have subsequently sparked debate, from criticisms that some of the indices do not capture robustness effectively (Veksler 2015), to those that have examined the impact of lobbying regulation on lobbying styles (Woll 2012).

Each of the four indices has defined how a lobbying law and its robustness can be conceptualised; each develops measures or indicators; and each allows researchers to generate scores by performing textual analysis of the lobbying

[2] See for example, Opheim (1991), Lowery and Gray (1997), Greenwood and Thomas (1998), Greenwood (1998), Thomas (1998), Yishai (1998), Rechtman and Larsen-Ledet (1998), Jordan (1998), Newmark (2005), Ozymy (2010, 2013), Chari et al. (2010), Holman and Luneburg (2012), Greenwood and Dreger (2013), Holyoke (2015), Veksler (2015).

[3] The literature on lobbying regulation has produced two other measurements of robustness, which are, however, excluded from this examination for two reasons. The first is related to the fact that Brining et al. (1993) do not provide guidelines on the construction of their measurement (Lowery and Gray 1997, 146). Consequently, it is not possible include their index in this investigation. The second reason is related to the exclusion of the measurement developed by Hamm et al. (1994). The low impact of their *lobbying registration constraint index* on the literature about lobbying regulations suggests the secondary importance of this measurement. The author's index has not been adopted by subsequent investigations on lobbying regulations.

regulation law, assigning points for each item of the measures and then summing scores that can be normalised for comparative analysis. As such, their analysis offers fertile ground to contribute to the debate on measurement validity and reliability in social science.

By linking the analysis of the four main indices in the lobbying literature to the larger methodological debate, this article thus seeks to understand which of the studied measurements is most valid and reliable. This methodological exercise is important because it adds empirical and theoretical insights into the larger literature of comparative politics, which has examined measurement validity and choices about concepts. In addition, for more general scholars of public policy, analysing the performance of the different indices that seek to capture the robustness of lobbying regulations adds methodological insights into the tasks that policy analysts must do with regard to conceptualisation and measurement of key policy matters. To be clear, we are not seeking to find which is the "best" or "single" winner in terms of how the indices we study perform. Rather, we are seeking to provide scholars with a replicable methodological standard for the investigation of measurements in social sciences, which may also provide insights for the development of new indices beyond those studied here.

The present article is structured as follows: the first two sections consider which of the indices developed in the lobbying regulation literature is the most *valid*, or which "indicator plausibly measures the conceptual ideas it is intended to measure" (Seawright and Collier 2014, 114). In this case, we consider alternative procedures for addressing the overall idea of measurement validity by focussing on different types of validation. To this end, we start with calculating the level of robustness (as measured by each of the four indexes) of 13 regulated jurisdictions around the world and present normalised scores. The section then pays attention to convergent validation and evaluates the levels of similarity between the indices, asking whether the final scores "produced by alternative indicators … (are) empirically associated and thus convergent" (Adcock and Collier 2001, 540).

The second section, which turns to content validation, then considers what elements are included and excluded in the indicators (Adcock and Collier 2001, 538), and thus seeks to measure the adequacy of content of the different indicators. It does so by using Organisation for Economic Co-Operation and Development (OECD) principles on lobbying laws as a baseline that constitutes a "gold standard" against which the indices can be evaluated. We argue that performing content validation using an international best standard as a guide is a novel and useful way to assess indices designed to evaluate public policy.

In light of criticisms made in the first two sections on both convergent and content validity, the third section then considers which among the indices is more *reliable*. Although this term is one which the literature has defined as meaning "repeated applications of a measurement tool yield(ing) consistent results" (Seawright and Collier 2014, 114), we also borrow ideas from natural science and argue that the term should also more explicitly incorporate the idea of "reproducibility", which is defined as "the variation in measurements made on a subject under changing conditions" – whereby changing conditions refer to a different rater or observer (Bartlett and Frost 2008, 467). On the basis of a coding test that we perform, the main question asked in the third section is thus as follows: when 25 younger scholars score lobbying legislation using these indices, which index is the most reliable? The question also serves as a foundation for making a more nuanced distinction between "validity" and "reliability". The standard view in the literature is "that although validity and reliability are distinct, a measure should not be considered valid if it is not reliable" (Seawright and Collier 2014, 114). However, the evidence uncovered suggests the need to reconsider this idea because "reliability" should be seen as a function of "repeatability" and "reproducibility" (Bartlett and Frost 2008, 467). Considering our analysis on the previous sections focussing on both convergent and content validation, we argue that the most "valid" measurement may not necessarily be the most reliable in terms of its "reproducibility", and the most "reproducible" is not necessarily the most "valid".

Convergent validation

This section first reviews the main literature on measuring the robustness of lobbying laws. We investigate the indices by Opheim (1991), Newmark (2005), Chari et al. (2010) and Holman and Luneburg (2012), retracing the authors' contribution to the interest group literature. The second part presents the result of a quantitative test performed on the coding applied to the regulations of Austria (which established its law in 2012), Australia (2008), Canada (2008), European Union (EU) (2011), France (2013), Germany (1951), Lithuania (2001), Mexico (2010), the Netherlands (2012), Poland (2006), Slovenia (2010), the United Kingdom (UK) (2014) and the US (2007). In order to evaluate the presence of dissimilarity between the four measurements under investigation, we present the results of the analysis of the standard deviations for the robustness scores of each case. This allows one to consider whether or not measures perform differently depending on the case of analysis. Next, we present the results of a correlation test.

Opheim's (1991, 405) pioneering work on measuring the rigour of lobby laws at the US state level represents the first contribution to the quantitative analysis of interest group regulation. She looks at 47 state regulations, identifying variations in terms of the degree of transparency and accountability that such laws guarantee. With the aim of investigating this variation, the author develops a measurement of the rigour of the regulation, which indicates the legislative independence and accountability from interest group pressure (Opheim 1991, 405). The measurement is based on a dichotomous coding procedure of the lobbying law according to 22 separately scored items drawn from three key dimensions of the lobbying regulation (Opheim 1991, 407). As the author argues, such key dimensions are "critical to the state's effort to regulate special interest activity" (Opheim 1991, 407). The three dimensions are as follows: the definition of a lobbyist (seven items), the frequency and quality of disclosure of personal and financial information (eight items) and the enforcement of the regulation (seven items). Each item is coded 1 if the regulation includes the item and coded 0 otherwise. The result is an additive index that scores from 0 to 22, where higher scores indicate more robustness. The impact of her work is reflected in the main contributions to the US interest group literature – namely, Brining et al. (1993), Hamm et al. (1994) and Lowery and Gray (1997) – either discussing or drawing upon this study. With the aim of producing a measurement, which is replicable over time, Newmark (2005, 185) revised Opheim's measure 15 years later. Newmark's robustness measure is based on 18 items, which include elements of how lobbying is defined in the regulation, what information lobbyists have to disclose and what activities pursued by lobbyists are prohibited by the law. The coding procedure is dichotomous and the additive index results in a measure that varies from 0 to 18. Similar to Opheim's index, high scores indicate high robustness. Unlike Opheim, Newmark's index does not include items on the enforcement of lobbying laws.

Cognisant of the upsurge of jurisdictions beyond the US that were pursuing lobbying laws in the 2000s, as noted by several scholars examining the worldwide experience,[4] Chari et al. performed a global comparative analysis of the robustness of lobbying laws. In order to do so, they applied an index developed by the American think tank *Centre for Public Integrity* to existing regulations.[5] This index, named the Centre for Public Integrity (CPI) index, results from a coding procedure based on 48 items and eight key elements of the regulations. These key elements are as follows: (1) the definition of lobbyists, (2) individual registration requirements, (3) individual disclosure of financial information, (4) employer spending disclosure, (5) electronic

[4] See, for example, Greenwood (1998), Thomas (1998), Rush (1998), Jordan (1998), Warhurst (1998), Ronit and Schneider (1998), Rechtman and Larsen-Ledet (1998), Hrebenar et al. (1998), Yishai (1998).

[5] This is done by applying the methodology of the hired guns: http://www.publicintegrity.org/2003/05/15/5914/methodology, last accessed on 14 October 2015.

filing, (6) public access to a registry of lobbyists, (7) enforcement and (8) revolving door provisions. In the case of the CPI index, the coding procedure is not dichotomous. The procedure weighs some items more than others, depending on whether the item is to be considered a critical feature of the key element.[6] The additive index results in a measure ranging from 0 (meaning low robustness) to 100 (meaning highest robustness). According to the levels scored by each regulation on the key elements of the CPI index, the authors distinguish between lowly regulated, medium-regulated and highly regulated systems (Chari et al. 2010, Chapter 4). At first glance, this index presented at least one advantage compared with the two previous measurements: its increased number of key elements (eight compared with three of Ophiem's measure) and items (48 compared with 22 and 18 of Opheim's and Newmark's indices) allow one to consider robustness more precisely across and within the elements of the regulation.

Finally, Holman and Luneburg (2012) were the first to consider the relative strictness of lobbying rules within European political systems. In their study, they provide a theoretical classification of regulated systems in line with Chari et al.'s contribution. The authors design their classification on 21 items that characterise lobbying regulation, which allows for an additive index to be made, as has been recently performed by Crepaz (2016a). Similar to previous studies, these items include the definition of lobbying, the disclosure requirements and enforcement of the rules. In addition, some items include whether the regulation is mandatory or voluntary,[7] whether or not the rules include the presence of codes of conduct for lobbyists and whether or not some interest groups are exempt from the rules. The regulated systems in Europe can be classified into strong or weak regulations depending on how they perform on the 21 items (Holman and Luneburg 2012, 21). Similar to Crepaz (2016a), we summed up the scores into a dichotomous index ranging from 0 (meaning low robustness) to 21 (meaning maximum robustness), with the aim of transforming the indicators provided by the authors into a quantitative measurement of the robustness of lobbying regulations. Compared with the previous measurements, this index presents the advantage of including features that are based on European lobbying regulations. However, its limited number of items does not allow one to consider all aspects of robustness in detail.

[6] For example, the procedure assigns a minimum score of 0 and a maximum score of 3 to item 11 (*is a lobbyist required to file a spending report?*). Item 15 on 'whether the spending in such report needs to be itemized or not' assigns a minimum score of 0 and a maximum score of 1 (as the answer to item 15 depends on the answer to item 11).

[7] Some European regulations are based on voluntary rules (EU or Germany) rather than mandatory. Such feature reduces the robustness of the regulation. The characteristics of such voluntary regulations are better explained in OECD (2012) or Crepaz and Chari (2014).

Table 1. Scores of 13 lobbying laws applying four different measurements of robustness

	Opheim (range of scores 0–22)	Newmark (0–18)	Chari et al. (0–100)	Holman and Luneburg (0–21)
Austria	9	11	32	17
Australia	4	3	33	8
Canada	14	9	50	15
EU	9	9	31	14
France	8	11	30	13
Germany	4	5	17	5
Lithuania	10	7	44	13
Mexico	6	7	29	13
The Netherlands	5	7	24	12
Poland	7	5	27	11
Slovenia	13	11	45	14
UK	9	7	27	8
US	12	11	62	18

Note: EU = European Union; UK = United Kingdom; US = United States.
Source: Own calculations.

With the above in mind, the robustness levels for 13 lobbying laws have been calculated applying the four methods of measurement. The results of such coding are shown in Table 1, which shows nominal scores. Our coding matches with the results presented by Chari et al. on Australia, Canada, Germany, Lithuania, Poland and the US. Similarly, our results are consistent with Holman and Luneburg (2012, 21) on Germany, Lithuania, Poland and Slovenia.

With the aim of allowing comparison of relative scores, the robustness scores have been normalised, ranging from 0 (meaning lowest robustness) to 1 (meaning highest robustness), as seen in Table 2.[8]

The results shown in Table 2 suggest that the robustness levels are more dissimilar for some cases than for others. Generally, the measurement by Holman and Luneburg scores higher values of robustness levels for all cases of analysis. On the contrary, the robustness scores using the measurements by Opheim, Newmark and Chari et al. tend to be more similar with the exception of some cases such as Austria and the EU. In the rest of the cases, at least two measures have equal or similar values. Interestingly, the calculated robustness score for the US shows a high similarity between the measures by Opheim, Newmark and Chari et al., suggesting that these

[8] Each total robustness score has been divided by its maximum (22 Opheim, 18 Newmark, 100 Chari *et al.* and 21 Holman and Luneburg).

Table 2. Normalised scores of 13 lobbying laws applying four measurements of robustness

Normalised values	Opheim	Newmark	Chari et al.	Holman and Luneburg	SD
Austria	0.41	0.61	0.32	0.81	0.22
Australia	0.18	0.17	0.33	0.38	0.11
Canada	0.64	0.50	0.50	0.71	0.11
EU	0.41	0.50	0.31	0.67	0.15
France	0.36	0.61	0.30	0.62	0.17
Germany	0.18	0.28	0.17	0.24	0.05
Lithuania	0.45	0.39	0.44	0.62	0.10
Mexico	0.27	0.39	0.29	0.62	0.16
The Netherlands	0.23	0.39	0.24	0.57	0.16
Poland	0.32	0.28	0.27	0.52	0.12
Slovenia	0.59	0.61	0.45	0.67	0.09
UK	0.41	0.39	0.27	0.38	0.07
US	0.55	0.61	0.62	0.86	0.14
SD	0.15	0.15	0.12	0.18	

Note: All scores vary from 0 to 1.
EU = European Union; UK = United Kingdom; US = United States.
Source: Own calculations.

measurements, developed on the American tradition of regulating lobbying, best apply to this case of analysis. The opposite is true for European lobbying laws, especially the Austrian law.

The last row of Table 2 shows the standard deviation on each measurement. Quantitative studies in political science typically have an approach that seeks to explain variation in the outcome (Mahoney and Goertz 2006). Hence, the more variation scholars have in the outcome, the better it is for their empirical investigation. In the case of this analysis, the measurement by Holman and Luneburg shows a higher variance compared with other measures. Figure 1 shows why this is the case.

The last box plot shows the maximum (0.86 for the US) and the minimum value (0.24 for Germany). The minimum is treated as an outlier in the distribution. This is because 92% of the scores vary between 0.38 and 0.86. The distribution of the measurements by Opheim and Chari et al. are symmetric and similar in terms of minimum, maximum, median and first and third percentiles. Newmark is dissimilar in its maxima (France, Austria and Slovenia with a score of 0.61). However, none of the scores takes higher values than 0.64. This confirms a main difference between the first three measurements and the last measure by Holman and Luneburg. The latter tends to score higher levels of robustness compared with the other measurements. The following considers this idea in more detail.

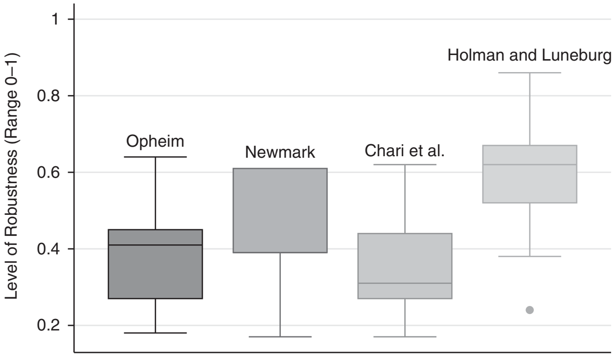

Figure 1 Distribution of the scores applying four measurements of robustness.
Source: Own calculations.

Low standard deviation scores indicate similarity in the robustness level using the four measures, whereas high scores indicate dissimilarity. For example, the high SD of Austria (0.22) suggests that the calculation of the robustness level using one measure or the other leads to dissimilar results. On the contrary, the opposite result applies to Lithuania, Slovenia, the UK (all SD < 0.10) and in particular Germany (SD < 0.05). In the next section, we will examine in more detail how one can account for the differences between the countries.

We will now evaluate whether the measurements under investigation perform differently on average by providing rank-order correlations between all measurements. This provides insights into the validity of each measurement using convergent validation (Adcock and Collier 2001; Rogers and Weller 2014). Correlations indicate the covariance of how the robustness scores are ranked, expressed as deviations from their mean. In other words, they indicate *average similarity*. The results are shown in Table 3.

The results of the correlation test suggest that there is a fairly close similarity on average between the four measures of robustness, although some findings are worth highlighting. The CPI index by Chari et al. correlates closely with Opheim's and Holman and Luneburg's measurements (0.74 and 0.73, respectively). Its correlation with Newmark's index is, however, low (0.51). The second lowest correlation is between Newmark's index and Opheim's index (0.68). Holman and Luneburg's index fairly correlates with all measurements under investigation (0.77, 0.81 and 0.73). This suggests that the American and European traditions of coding lobbying laws are not a dividing principle. From this test of convergent validation, we conclude that, on average, Holman and Luneburg's measurement is most *similar* to other measurements, whereas Newman's is most *dissimilar*.

Table 3. Rank-order correlations between the scores of 13 lobbying laws using four measures of robustness

Correlations	Opheim	Newmark	Chari et al.	Holman and Luneburg
Opheim	1	0.68	0.74	0.77
Newmark	0.68	1	0.51	0.81
Chari et al.	0.74	0.51	1	0.73
Holman and Luneburg	0.77	0.81	0.73	1

Source: Own calculations.

In conclusion, the four measurements of robustness perform differently depending on the case of analysis: evidence has shown that some indices disagree on the robustness levels of some regulations. This is particularly the case of the measurement by Holman and Luneburg. Compared with the other measures, this index tends to score higher levels of robustness. The other indices by Opheim, Newmark and Chari et al. show a higher degree of similarity, especially when it comes to coding the US legislation. However, on average, Newmark's index shows a higher degree of dissimilarity compared with the other measurements, whereas Holman and Luneburg's index shows the highest similarity.

A limitation to convergent validation, however, is that high correlations among indicators "may reflect factors other than valid measurement. For example, two indicators may be strongly correlated because they measure some other concept; or they measure different concepts, one which causes the other" (Adcock and Collier 2001, 541). To dispel any possibility of misspecification between concept and measurement, it is necessary to consider another method of validation, which seeks to examine whether an indicator "capture(s) the full content of the systematised concept" (Adcock and Collier 2001, 541), or, namely, content validation.

Content validation

This second section investigates and evaluates the indices' construction, probing whether the items of the different measurements consider all relevant conceptual aspects of robustness of the lobbying regulation. Adcock and Collier (2001, 539) examine various examples of content validation by focussing on developments in the literature on democracy, highlighting how authors such as Paxton (2000) consider the problems of "omission of key elements from the indicator and inclusion of inappropriate elements".

Although not developed in the extant literature, a novel way to evaluate what may be missing in an indicator, which is particularly useful when examining indices that attempt to evaluate a public policy, is to consider a relatively objective "gold standard" from a reputable international organisation that outlines what should be entailed in a law in terms of best practice, and then see how well indices capture this. Such benchmarks are established by various international organisations such as the OECD and the Council of Europe for various public policies pursued at the domestic level, such as lobbying regulation, whistleblowers legislation, ethics reform laws and privacy protection laws to name a few.[9]

One reason for taking the OECD standard as a point of comparison is that national policymakers often seek inspiration in norms established by international organisations. Previous studies have shown that national policymakers gain such inspiration from standards that are established by international organisations (True and Mintrom 2001; Stone 2004). Yet, we should be careful to not consider the policy recommendations of international organisations as being scientific standards for validation, as these policy positions may be the result of subjective deliberations, or simply ill-informed discussions, which may have left out important indicators of what constitute robust lobbying laws. However, the relevance of the policy recommendations of international organisations about the adoption of lobbying laws has been increasingly acknowledged in the academic literature and in the real world of politics. For instance, Crepaz (2016b) found that the OECD and the EU recommendations about lobbying regulations have been successful in encouraging the adoption of lobbying laws among member states. This finding is supported by qualitative evidence: the Austrian and Irish governments introduced their lobbying law following the principles established by the OECD.[10] With this justification in mind, we decide to conduct content validation considering the elements that characterise robust regulations according to the OECD's (2008) report on lobbying regulations, which is worth considering in some detail.

[9] Beyond the OECD (discussed in this section), the Council of Europe is in the process of establishing key recommendations that states should follow when developing lobbying laws (see http://www.coe.int/t/dghl/standardsetting/cdcj/Lobbying/Lobbying_en.asp). Key recommendations from the Council of Europe regarding best practices when states develop Whistleblowing legislation can be seen on http://www.coe.int/t/dghl/standardsetting/cdcj/Whistleblowers/protecting_whistleblowers_en.asp. On key OECD principles for managing ethics in the public service, see http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=129&Lang=en. Moreover, on privacy protection, see the OECD's principles on http://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm, last accessed on 15 June 2016.

[10] See http://www.parlament.gv.at/PAKT/VHG/XXIV/ME/ME_00293/imfname_223599.pdf for Austria and http://www.per.gov.ie/en/regulation-of-lobbying/ for Ireland, last accessed on 8 May 2016.

The OECD has focussed on the promotion of lobbying laws over the last decade through its Department of Government Integrity, dedicating particular attention to the topic of regulating lobbying as part of fighting corruption in the public sector. In 2008, the OECD published a report containing guidelines for the introduction of robust lobbying rules in its 34 member states. Such guidelines are based on the conceptual elements of *defining lobbying*, *disclosure requirements*, *reporting processes and technology*, *timeliness* (namely update of information), *enforcement and compliance* (OECD 2008, Chapter 2). According to the report, regulations that extensively consider these five key elements are to be considered robust (OECD 2008, 16). Hence, the more the construction of the indices under investigation captures these key items, the more such measurements catch the concept of robustness as defined by the OECD, which thus represents what can be deemed to be the "gold standard".

The key items, and how the different indices capture them, are as follows:

1. *Defining lobbying*: in order to be robust, lobbying laws need to clearly define lobbying, critical for understanding "who is to be regulated" (OECD 2008, 42). There are three main aspects to defining lobbying. First, the regulation needs to distinguish between government officials and lobbyists (OECD 2008, 42). The measurements by Opheim and Newmark include this in items 3 and 4, asking whether the definition of lobbyists considers public officeholders as well. Items 5, 6 and 7 of the same measurements consider whether a minimum compensation, expenditure or time standard apply to the definition of lobbyist. Similarly, item 2 of the CPI index by Chari et al. include the same aspects to the definition of lobbying activity. Second, a robust definition of lobbying should also include all interest group categories, meaning that it should aim at regulating consultant lobbyists acting on behalf of third parties, in-house corporate lobbyists and NGOs (OECD 2008, 44). Only the index by Holman and Luneburg considers whether the regulation affects lobbying consultancies, for-profit interest groups and nonprofit interest groups equally (items 3, 4 and 5). Third, the definition of lobbying needs to clearly identify state actors targeted (OECD 2008, 46). The measures by Opheim and Newmark consider whether the regulation affects lobbyists seeking contacts with members of Parliament and civil servants (items 1 and 2). Ministers are therefore excluded from these items. On the contrary, Holman and Luneburg include both legislative and executive lobbying in items 6 and 7. Finally, the CPI index assumes that legislative lobbying is covered and codes executive lobbying only (item 1).
2. *Disclosure requirements*: a further critical aspect of regulating lobbying is related to the disclosure requirements. Typically, lobbyists need to enter a certain amount of information in a (public) register before establishing

any contact with public officeholders, where the amount and accuracy of the information is directly related to the law's robustness. This includes personal details, objectives of lobbying and financial disclosure (OECD 2008, 50, 51, 57, 58). Regarding the construction of the measurements, only the indices by Chari et al. and Holman and Luneburg consider the disclosure of personal information (items 8, 9 and 10 for both indices). These include personal details such as name, address, contacts and photograph. When it comes to items concerning the disclosure of objectives of lobbying, again Chari et al. and Holman and Luneburg provide the most complete index construction. Item 5 of the former index and items 11 and 12 of the latter consider whether lobbyists need to disclose the subject matter of the lobbying activity or even who is targeted. Another item in Opheim's index (item 12) asks whether interest groups approve or oppose the legislation they are seeking to lobby on at the moment of registration.[11] On the disclosure of financial information, all four measurements have some items that consider total spending, categorised spending, total income of lobbyists and sources of income. In addition, only the CPI index includes separate items for campaign contributions and gifts (items 23 and 24). The same measurement also considers whether employers of lobbyists need to submit financial information (items 26 and 27). On the contrary, Holman and Luneburg do not consider gifts and other forms of donations benefitting public officeholders. Newmark focusses on whether campaign contributions and donations are considered as prohibited activities rather than to be disclosed (items 15–18). The last section of this key element focusses on *other* disclosure requirements (OECD 2008, 60), where additional elements are found as follows: Opheim's index provides one item on the disclosure of potential conflicts of interest or influence peddling (item 15); Holman and Luneburg consider whether lobbyists need to disclose every contact with public officeholders (item 17); and Chari et al. collect any additional form of disclosure (item 10.)

3. *Timeliness and ethics*: the OECD report (2008, 66) evaluates the quality of disclosure, such as the frequency of registration and updates of the information in the register. Opheim and Newmark provide items on the frequency of registration (items 8). Chari et al., more in detail, provide items on whether registration needs to be completed before lobbying (items 4 and 6), whether the regulation sets rules on the

[11] We know that the activity of lobbying is more complex than simply agreeing or disagreeing with a piece of legislation, but Opheim's item helps to distinguish *advisory* lobbying from *confrontational* lobbying aimed at opposing the introduction of certain pieces of legislation and their legitimacy.

notification of changes in the registered information (item 7), whether lobbyists need to submit a no activity report (item 25). Further, the OECD (2008, 67) outlines that lobbyists should sign a code of conduct, where Holman and Luneburg uniquely consider such codes of conduct in item 21.

4. *Reporting processes and technology*: this element focusses on the electronic filing of information and the public access to such information (OECD 2008, 65–66). Although Opheim and Newmark do not dedicate items to this, Holman and Luneburg consider whether or not the information on the lobbying register is accessible by the public (item 20).[12] Chari et al. dedicate more attention to this aspect by considering whether or not the access to the register is public (items 31–34), the register provides users with summary reports (items 35–37), lists are updated frequently (item 38) and lobbyists can file their reports electronically (items 28, 29).
5. *Enforcement and compliance*: the *first* of four dimensions of this element (OECD 2008, 71–73) considers education and training on lobbying regulation compliance. Chari et al.'s index uniquely considers this aspect (item 30). The *second* aspect focusses on whether the regulation is voluntary or mandatory in nature. Here, only Holman and Luneburg include this in their robustness measure (item 1). The *third* aspect considers the enforcement of the rules (and the efficiency thereof) by the public body entrusted to monitor compliance. On the power of the monitoring agency, Opheim's includes its ability to review information and reports (item 16), demand subpoena witness or record (items 17, 18), conduct administrative hearings (item 19), apply fines or penalties in cases of noncompliance (items 20, 21) and file an independent court action (item 22). Although Newmark does not include any items on this key element, Chari et al.'s index has a high number dedicated to this, including the power to review reports (item 40), publish the names of lobbyists infringing the rules (item 47), apply fines or penalties for noncompliance (items 41, 42, 44, 45) and ask about the last levied case of noncompliance (items 43, 46). Less detailed, Holman and Luneburg include whether the monitoring agency can apply fines or penalties for cases of noncompliance in one single item (19). The *fourth* aspect of this key element concerns cooling-off periods, which delay politicians and civil servants from entering the lobbying industry. Of all measurements, only the CPI index used by Chari et al. considers this (item 48).

[12] Opheim does not consider items on this key element because of the absence of E-Government services at the time of writing.

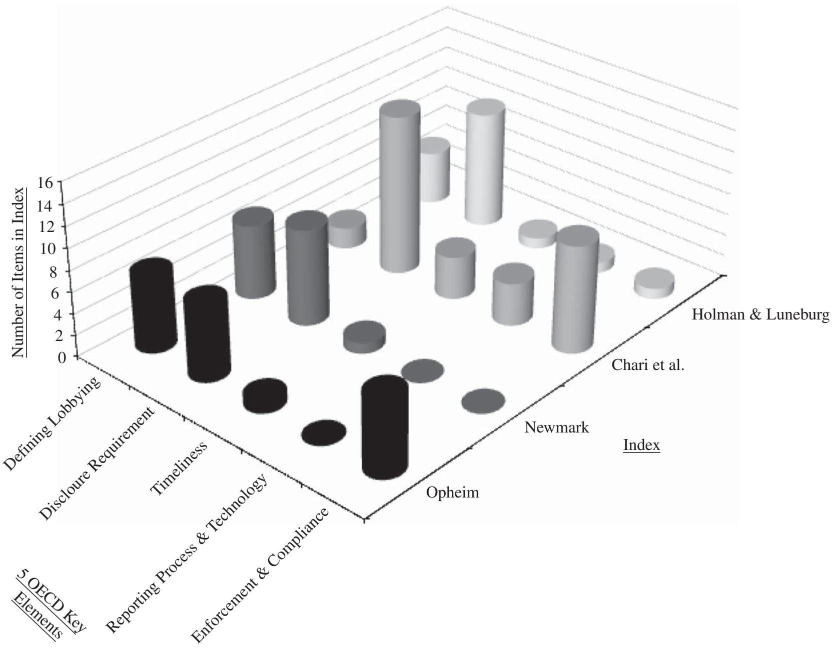

Figure 2 Number of items (per measurement) that fall in the five Organisation for Economic Co-Operation and Development key elements of the robustness of lobbying laws.
Source: Own calculations.

In Figure 2, we illustrate the number of items (per measurement) that fall within the conceptual dimensions of the robustness of lobbying laws as identified in the OECD policy recommendations. The figure shows, for example, that, on the first key element *Defining Lobbying*, the measurements by Opheim and Newmark count seven items each, whereas Chari et al.'s and Holman and Luneburg's indices dedicate two and five items, respectively, to this dimension. This suggests that, compared with others, the first two measurements demonstrate more content validity on this element, meaning that the index is better able to catch the concept of robustness using the OECD's report as *gold standard*. Considering all five OECD's key elements of the robustness of lobbying laws, Chari et al.'s index appears as the most valid (in terms of content) as the highest number of items falls in conceptual dimensions identified by the OECD, followed by Opheim's index, Holman and Luneburg's and then Newmark's.

It is interesting to note that using the OECD's conceptual dimensions also allows us to better understand the differences in the robustness levels across the countries, which was seen in the previous section on convergent validation.

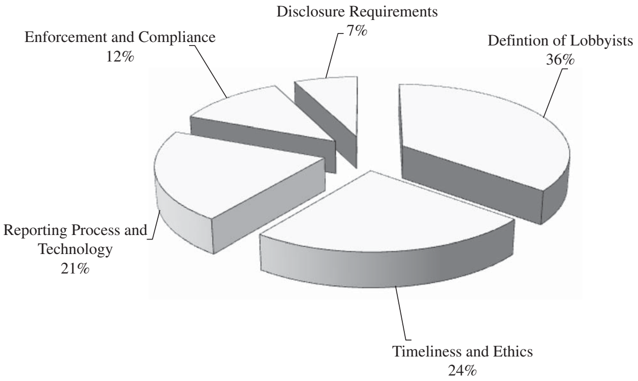

Figure 3 Percentage of dissimilarity between measurements by conceptual dimensions of the robustness of lobbying laws (averages calculated on all 13 pieces of legislation). *Source*: Own calculations.

In terms of method, we analyse the levels of robustness by disaggregating the total scores for each of the indices for the 13 pieces of legislation into the five conceptual dimensions of the OECD. We assign each item of the indices to one of these conceptual dimensions (as seen in Figure 2) and calculate the mean robustness level of the dimension for each country. This results in 20 observations for each country (260 observations in total). Next, we calculate the average standard deviation for each dimension considering all 13 pieces of legislation. This process results in five means (one for each dimension of the OECD) that indicate levels of dissimilarity between measurements. High averages indicate dissimilarity between the measurements concerning a given conceptual dimension, whereas low scores signify similarity. Figure 3 displays these levels of similarity for each conceptual dimension in percentages.

The results suggest that the four measurements are the most dissimilar in the conceptual dimensions of *definition of lobbyists* (36%) and (to a lesser extent) *timeliness and ethics* (24%). From this perspective, a dividing principle between the measurements is the way in which they score "how lobbying is defined by the regulation" and "how frequently information is disclosed and updated". Together, these two dimensions account for 60% of the total dissimilarity between measurements. The levels of similarity between the scores of the items assigned to the dimensions concerning disclosure requirements, the reporting processes and the enforcement of regulation, on the contrary, reveal, on average, a higher degree of similarity. From these observations, we can conclude that the way indices capture the dimensions of the *definition of lobbyists,* in particular, influences the calculation of the robustness level and that the indices under consideration measure the latent concept on, at least, four out of five of the OECD's conceptual dimensions.

Summarising the main findings for this section, the results in favour of the CPI index suggest that, if the right key elements are addressed in the construction of the index, having more items makes an index more valid. In other words, if you ask more questions, you are more likely to come up with a more complete and objective answer. In fact, Opheim, which is constructed by 22 items, scores second, Holman and Luneburg, which is constructed by 21 items, scores third and Newmark with its 18 items scores last.

Nevertheless, content validation is limited by a trade-off between parsimony and completeness that arises because indicators routinely fail to capture the full content of a systemised concept. Capturing this content may require a complex indicator that is hard to use and adds greatly to the time and cost of completing the research. It is a matter of judgement for scholars to decide when efforts to further improve the adequacy of content may become counterproductive (Adcock and Collier 2001, 539).

The next section thus elaborates on this argument in more detail, first considering how the trade-off referred to impacts on reliability and then focussing on the indices.

Reproducibility

In the previous sections on methods of validation, we concluded that Holman and Luneburg's and Chari et al.'s indices appear to be the most valid (the former in terms of convergence and the latter in terms of content). From these observations, we inferred that high-dimensional indices tend to be more valid than lower-dimensional ones. However, we argue that validation is not the only criteria upon which to judge a measure as it might favour the construction of complex and high-dimensional indices. In this section, we introduce a third criterion of evaluation – namely, *reliability* – which indicates whether a measurement technique secures consistent results upon repeated application. We argue that reliability is a function of "repeatability" and "reproducibility". Both repeatability and reproducibility refer to a variation in the level of agreement adopting the same measurements on the same subject. However, repeatability is attained when the same observer or rater attains consistent measurements, whereas reproducibility is attained when different observers or raters attain similar results to the initial observer (Bartlett and Frost 2008, 467–468).

The importance of the "lack of reproducibility" is based on an emerging concern within the natural science community related to challenges in reproducible research as recently examined in a series of special reports by *Nature*.[13] As an example in natural science, let us say Medicinal Chemist "A" has developed procedures to obtain a significant yield for a new compound, which may be of value not only to fellow academics, but also to the pharmaceutical industry that may use the compound to help develop a new drug. However, the pharmaceutical company may think twice about using A's procedures if they are long, complex and have a low reproducibility rate. As such, for practical reasons, the company may consider a simpler, more parsimonious, procedure developed by Medicinal Chemist B to develop the same or similar structure, even if it gives a lower yield.

Analogously, one may argue that the rise of lobbying laws over the last 15 years means that many governments throughout the world are relying on scholars' work on lobbying regulation when developing their laws, as evidenced in various submissions, committee reports and proceedings from the UK, Scotland and Ireland.[14]

Governments may also consume this research when developing their own laws as well, not dissimilar to how a pharmaceutical company may consume research from a medicinal chemist. Let us say that a team of civil servants and their relevant Minister are tasked to develop a draft lobbying bill. After they have developed this draft, the team may seek to measure what "level of robustness" the potential law will have before it is presented to Cabinet. This would be done in order to better gauge how much transparency and accountability the new potential law provides for. The importance of doing this can be illustrated with two hypothetical scenarios. If it is found that the draft bill only provides for a lowly regulated environment, then this may be unacceptable for a newly elected government that wants to "clean up" politics in a country historically riddled with corruption. Conversely, a government of a country with low levels of corruption may prefer to develop less robust legislation given higher costs associated with developing tighter regulation. In both scenarios, the policymaking team that has measured the robustness of the draft bill may potentially make changes before final presentation to Cabinet. For example, relying on Chari et al., if the policymakers' analysis of the draft bill reveals a point score of more than 60 points, then this would be representative of law in a highly regulated system, between 30 and 59 points, medium regulation, and between 0 and 29, low regulation.

[13] See http://www.nature.com/news/reproducibility-1.17552, last accessed on 12 December 2015.

[14] On Ireland, see http://www.per.gov.ie/en/regulation-of-lobbying/, on Scotland http://www.scottish.parliament.uk/S4_StandardsProceduresandPublicAppointmentsCommittee/Inquiries/evidence_summary_for_web.pdf, on the UK https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/407530/2901376_LobbyingStandards_WEB.pdf, last accessed on 12 December 2015.

With this in mind, the team of policymakers would want to know which index can be applied simply and reliably, wherein the attained results are the same as researchers in the field who apply the index for their own academic work: "which of the four indexes gives me a reliable score?" That is, the team would want an index that does not suffer from a lack of reproducibility because it is so complex to apply. Extending on this logic, one may even argue that interest groups that feed into the consultation process when lobbying laws are developed may equally want to know which index is the most reliable when seeking to measure the robustness of lobbying bills that are eventually tabled by Cabinet. Equally, scholars who are not necessarily experts of lobbying regulation may want to know which is the most reliable index to measure the robustness of lobbying laws of a country – and then compare this with, for example, the strength of FoI legislation – in order to better evaluate the "sunshine laws" of said state.

In this regard, having a more complex indicator when evaluating public policy is particularly important not only in terms of costs to researchers but also to practitioners who seek to reproduce results. It is reasonable to hypothesise that more items in an index could lead to a higher complexity of the coding procedure, and therefore increase the error when trying to measure the robustness of the law. With this in mind, this section considers whether or not the most valid indicators as seen in the previous sections are really the ones that scholars and practitioners would want to use in the context of being reliable, which refers to minimising measurement error.

This reliability test is based on the comparison between our results presented in Table 2 in the first section and the results produced by trained coders. It is important to note that coding is a common procedure when research in social science involves the analysis of texts, where such procedures have been developed in measuring the position of political parties on a left–right dimension (Mikhaylov et al. 2012), the degrees of populism (Rooduijn and Pauwels 2011) and research on interest groups (Boräng et al. 2014). However, when human coding is involved, vague, misleading or incomplete coding procedures may lead to arbitrariness in the coding (Krippendorff 2004). Reliable coding procedures must not leave space for interpretation to coders (Mikhaylov et al. 2012). In order to be perfectly reliable, a coding procedure needs to lead to the same result regardless of the coder. This is very unlikely and developers of coding procedures normally aim at minimising *intercoder error*.

Our coding results were produced in two stages. First, a pilot-coding test was performed on 10 coders. This allowed us to collect the robustness scores on 10 out of the 13 pieces of legislation studied in the article: Austria, Australia, the EU, France, Germany, Lithuania, the Netherlands, Poland, UK and the US (2007).[15] This preliminary test resulted in 40 robustness scores, which were compared with our coding results (seen in the first section). This pilot stage also saw that the EU legislation registered the lowest level of reliability. This resulted in our decision to focus on the EU with a larger sample of coders in the second stage in order to guarantee a higher degree of accuracy.

The second coding test thus involved 15 trained coders that focussed exclusively on EU lobbying regulations and used all four indices. This resulted in 60 robustness scores on one piece of legislation, reducing the confounding factor of coding different pieces of legislation. According to Benoit et al. (2016), the number of coders is to be considered adequate, as standard deviations tend to not vary when such tests involve 15 coders or more. The coders were familiar with the EU lobbying law of 2011 and were familiar with the literature on lobbying regulation as well as the coding procedures found therein. The coders were third-year undergraduate students enrolled in the module on EU politics, which also covers the issue of lobbying in the EU and policies aimed at increasing transparency in the European Parliament and in the Commission. One traditional career path of these students is to work in the civil service, meaning that they are representative of future policymakers who may one day be involved in developing transparency laws. The students chose to partake in the test on a voluntary nature given their interest in the topic. The Ethics Research Committee of the School of Social Sciences and Philosophy, Trinity College Dublin, approved the methodology and the procedures of the test.

In terms of the findings, we first present the distribution of the coding error, where the error is expressed in proportion of agreement between coders. For each code assigned to each item of the measurements' construction, we calculated the level of agreement (range 0–1), as displayed in Figure 4. The closer the level of agreement for each item is to 1, the better the index performs in terms of reliability on that particular item. The *y*-axis shows the proportion of levels of agreements. For instance, a value of 0.8 on item 10 of Opheim's index means that 80% of the coders answered in the same way to item 10 of Opheim's coding methodology.

In the distribution of the levels of agreement between coders for Chari et al.'s index, we also included the number of possible answers to the items (according to the scoring methodology). For example, the scoring methodology for item 1 is dichotomous, whereas for item 2 coders can

[15] The coding was done by PhD students in Political Science. Each coder coded one piece of legislation by applying each of the four measurements under investigation. For each of the four measurements, the surveys, the instructions to coding, and all relevant documents of the law were given to the coders.

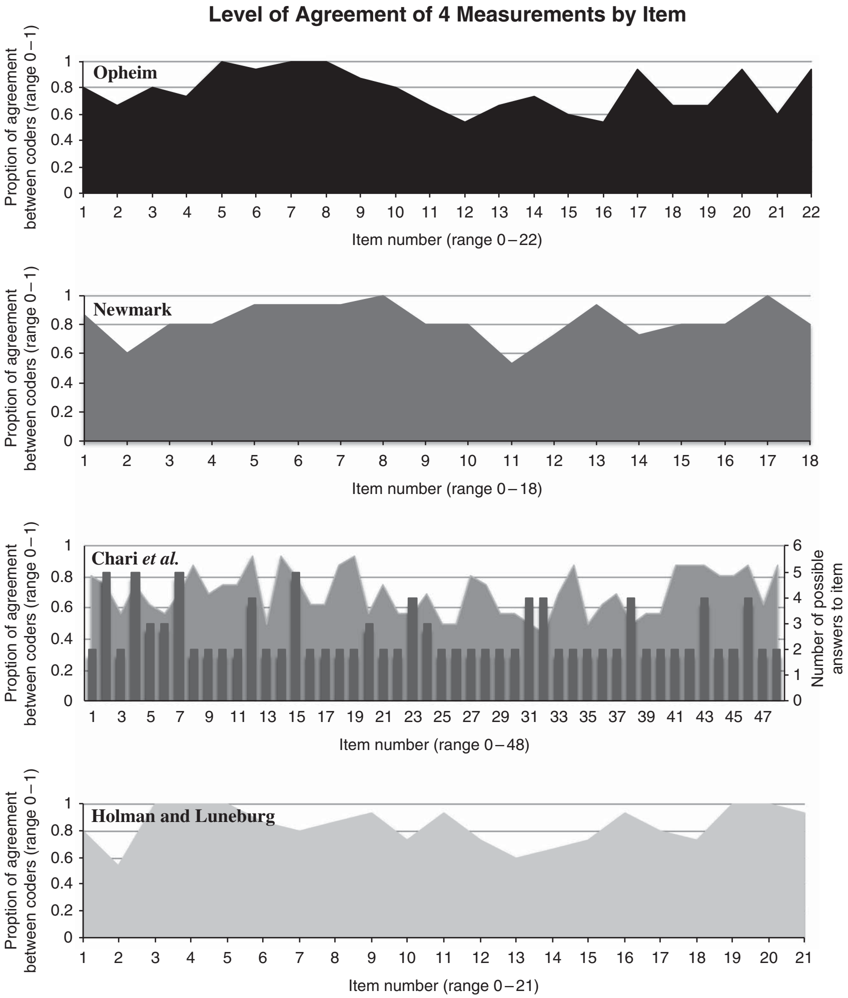

Figure 4 Distribution of the coding error expressed in proportion of agreement for each item.
Source: Own calculations.

assign a score choosing from five different options. We account for the number of possible answers because we believe that a more complex answer scheme could impact the overall level of reliability of the index.[16] For example, the proportion of agreement of Chari et al.'s index for

[16] We do not account for this factor using Opheim's, Newmark's and Holman and Luneburg's indices because their coding methodology relies on a dichotomous answer scheme.

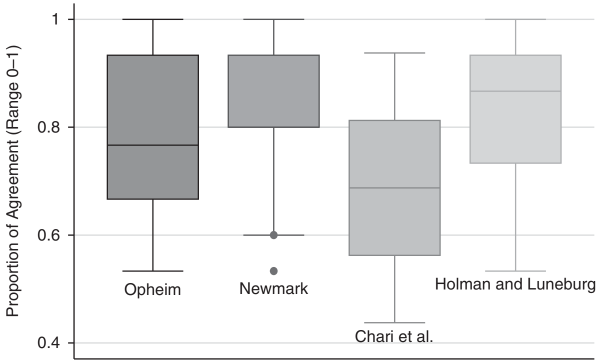

Figure 5 Distribution of the levels of agreement for each measurement.
Source: Own calculations.

item 32 is below 0.50. This might be due to the fact that the scoring methodology allows coders to choose between four different answers to item 32 with the risk of reducing the overall reliability of the index on this particular item.

The results shown in Figure 4 suggest that the index by Chari et al. appears to have the lowest intercoder agreement expressed in proportion of agreement for each item. The measurements by Opheim, Newmark and Holman and Luneburg show higher levels of agreement. In particular, the index by Holman and Luneburg reaches levels of 100% agreement in 24% of the items and agreement levels above 90% in 43% of the items. This is the highest result compared with the other measurements. On average (for all items and for each measure), the proportion of agreement equals 70% for the index of Chari et al., 78% for Opheim, 82% for Newmark and 84% for Holman and Luneburg. This is better shown in Figure 5, which illustrates the distribution of the levels of agreement for each measurement.

The results shown in Figure 5 suggest that levels of agreement, on average, seem to be highest for the index by Holman and Luneburg and lowest for the CPI index of Chari et al. To test whether this agreement is statistically consistent, we provide the results of an intercoder agreement test using *Krippendorff's α* statistic (as shown in Table 4), which is commonly used by researchers to evaluate levels of agreement (Mikhaylov et al. 2012).

All calculated *α*s are statistically significant, meaning that levels of agreement are not random (Lombard et al., 2002). The level of agreement varies from fair (for the indices of Opheim and Chari et al.) to moderate (for the indices of Newmark and Holman and Luneburg). Above all, the index by Holman and Luneburg has the highest *α statistic*, confirming the argument that levels of agreement are more consistent throughout all elements that

Table 4. Intercoder agreement using Krippendorff's α statistic

Measurements	α Statistic	p-Value	Agreement
Opheim	0.35	0.00	Fair
Newmark	0.43	0.00	Moderate
Chari et al.	0.32	0.00	Fair
Holman and Luneburg	0.46	0.00	Moderate

Source: Own calculations. Level of Agreement is defined by the categories of no agreement (0), slight agreement (0.01–0.20), fair agreement (0.21–0.40), moderate agreement (0.41–0.60), substantial agreement (0.61–0.80) and almost perfect agreement (0.81–1) (Landis and Koch 1977).

compose the measurement. One may reasonably argue, however, that there is no measure that is more reliable than "moderate", which is still not very high. This lack of reliability for scholars interested in lobbying regulations suggests that, although some perform reasonably well, there is room for improvement to develop new ones in future research. A unique way of evaluating whether or not a new index is reliable would be to fruitfully test its reliability throughout the development stage of the index, asking coders to score legislation, as was performed in this section.

Conclusions

The present article's goal was to provide researchers with methodological insights for the assessment of validity and reliability of measurements in the field of public policy and, in particular, of the indices used to evaluate policy outputs. To do so, we comparatively examined 13 jurisdictions worldwide in order to assess the robustness of lobbying laws using four established indices that measure the levels of transparency and accountability in this regulatory policy. We argued that a solid way of evaluating the performance of a measurement in this field of research is to apply three criteria of judgement, which are linked to the larger, methodologically based literature. First, we conducted convergent validation to assess to which extent measurements perform differently depending on the case of analysis. Second, we performed content validation with the aim of identifying which conceptual dimensions are best captured by the different measurements. Finally, we explored levels of reliability by paying particular attention to the issue of reproducibility (namely, the level of agreement between different coders in the application of the measurement). The main methodological, theoretical and empirical insights of value to the larger field, especially those seeking to better understand validity and reliability of measurements when evaluating public policy, are threefold.

The first significant finding stems from the second section of the article that focussed on content validation. From a *methodological* perspective, we argued that a novel way to evaluate what is missing in an indicator, which is useful when examining measurements that attempt to evaluate a public policy, is to consider a relatively objective "gold standard" from international organisations. Reputable organisations, such as the OECD and the Council of Europe, have established principles regarding what constitutes "best practice" in some key regulatory policy areas social scientists are trying to measure. Moving forward, when public policy scholars are seeking to better understand whether an indicator adequately captures the content of a concept, they may examine how such indices capture these key items outlined by such organisations.

Another significant insight stems from our discussion regarding which of the indices is more *reliable*, defined as giving consistent results that reproduce the index scores (as seen in the first section). From a *theoretical* perspective, this section has taken from ideas and concerns raised in natural science, arguing that reliability in social science should be seen as being more than simply arriving at consistent results. Rather, reliability, which can be seen as a function of both repeatability and reproducibility, should be understood in terms of providing consistent results that do not suffer from a lack of reproducibility. This is significant for researchers, who may seek to replicate investigation. It is also important for public policy practitioners who may effectively be "consumers" of academic research by taking academic measurements/indices and applying them to the development of legislation, as seen when practitioners use the indices studied in this article to help devise lobbying laws. If the scores arrived at by either academics or practitioners using the indices are different to what the experts would arrive at using the same index, then the measurement instrument suffers from a lack of reproducibility, regardless of whether or not it is the most valid.

In light of this, scholars may see this article as offering mixed results. The *empirical* investigation focussing on convergent validation highlighted the supremacy of Holman and Luneburg, whereas that focussing on content validation highlighted that of the CPI index used by Chari et al. In terms of reliability, the reproducibility test showed that Holman and Luneburg's index ranks the highest, even though none of the four measures is more reliable than "moderate". When comparing the results across all three sections, we effectively saw that the strength of Chari et al.'s CPI (in terms of content validity) was also its bane (in terms of its reliability): although asking an exhaustive set of questions may help capture robustness of regulatory legislation more completely, having a multitude of questions runs the risk of creating less consistent answers when this is reproduced by other coders.

However, any feelings of "mixed results" should be countered with an important lesson learnt for public policy scholars, representing the third significant insight from this study: the evidence suggests that the most "valid" measurement may not necessarily be the most reliable in terms of its "reproducibility", and the most "reproducible" is not necessarily the most "valid". This finding may be less than satisfactory in terms of arriving at an "overall winner" among the four measurements studied, which was not the goal of this study. However, the work serves as a basis for scholars of lobbying regulation to consider constructing new indices that are both the most valid *and* reliable. To this end, one solution is to include having coders constantly involved in the development of any future index in order to ensure its future reproducibility.

Without doubt, the way governments have regulated lobbyists has changed over the last 25 years, and so has the concept of *robustness*. Technological advancements and the growing complexity of lobbying environments have influenced the way governments seek to increase transparency and accountability in lobbying. Other relevant concepts of analysis, such as the *quality of democracy* or *varieties of capitalism*, might undergo similar conceptual evolutions. One the one hand, measurements therefore need to constantly adapt to new elements that define the systemised concept they seek to capture. On the other hand, researchers need to investigate these dimensions in relation to existing measurements with the final goal of producing consistent empirical research. With this article, we hope to have offered researchers methodological insights to help engage with these endeavours.

Acknowledgements

The authors thank the three anonymous reviewers as well as the editors for their constructive comments and suggestions on earlier versions. The authors are also indebted to the students of Trinity College Dublin's third-year undergraduate class in EU Politics of 2014–2015 who generously gave their time and partook in the coding experiment (as discussed in the section on reproducibility). Finally, M. C. acknowledges the support of Irish Research Council, which partly funded this study, and R. C. thanks Isabel Rozas for keen insights given during the research phase.

Supplementary material

To view supplementary material for this article, please visit https://doi.org/10.1017/S0143814X16000271

References

Adcock R. and Collier D. (2001) Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review* 95(3): 529–546.

Banisar D. (2006) *Freedom of Information Around the World 2006: A Global Survey of Access to Government Information Laws*. London, UK: Privacy International.

Bartlett J. W. and Frost C. (2008) Reliability, Repeatability and Reproducibility: Analysis of Measurement Errors in Continuous Variables. *Ultrasound in Obstetrics and Gynecology* 31(4): 466–475.

Benoit K., Conway D., Lauderdale B. E., Laver M. and Mikhaylov S. (2016) Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data. *American Political Science Review* 110(2): 279–285.

Boräng F., Eising R., Klüver H., Mahoney C., Naurin D., Rasch D. and Rozbicka P. (2014) Identifying Frames: A Comparison of Research Methods. *Interest Groups & Advocacy* 3(2): 188–201.

Bowman K., Lehoucq F. and Mahoney J. (2005) Measuring Political Democracy: Case Expertise, Data Adequacy, and Central America. *Comparative Political Studies* 38(8): 939–970.

Brining M., Holcombe R. and Schwartzstein L. (1993) The Regulation of Lobbyists. *Public Choice* 77(2): 377–384.

Chambers S. A. and Carver T. (eds.), (2008) *William E. Connolly: Democracy, Pluralism and Political Theory*. London, UK: Routledge.

Chari R., Hogan J. and Murphy G. (2010) *Regulating Lobbying: A Global Comparison*. Manchester, UK: Manchester University Press.

Clarke B. (1979) Eccentrically Contested Concepts. *British Journal of Political Science* 9(1): 122–126.

Collier D. F., Hidalgo D. and Maciuceanu O. A. (2006) Essentially Contested Concepts: Debates and Applications. *Journal of Political Ideologies* 11(3): 211–246.

Connolly W. E. (1974) *The Terms of Political Discourse*. Lexington, MA: Heath D.C. and Co.

Crepaz M. (2016a) Investigating the Robustness of Lobbying Laws: Evidence from the Austrian Case. *Interest Groups & Advocacy* 5(1): 5–24.

Crepaz M. (2016b) Why Do We Have Lobbying Rules? Investigating the Introduction of Lobbying Laws in EU and OECD Member States. Paper presented at the 87th SPSA Annual Conference in San Juan, January, 8th 2016, Puerto Rico.

Crepaz M. and Chari R. (2014) The EU's Initiatives to Regulate Lobbyists: Good or Bad Administration? *Cuadernos Europeos de Deusto* 51(1): 71–97.

Gallie W. B. (1955) Essentially Contested Concepts. *Proceedings from the Aristotelian Society* 56(1): 167–198.

Greenwood J. (1998) Regulating Lobbying in the European Union. *Parliamentary Affairs* 51(4): 587–599.

Greenwood J. and Dreger J. (2013) The Transparency Register: A European Vanguard of Strong Lobby Regulation. *Interest Groups & Advocacy* 2(2): 139–162.

Greenwood J. and Thomas C. S. (1998) Introduction: Regulating Lobbying in the Western World. *Parliamentary Affairs* 51(4): 487–488.

Hamm K., Weber A. and Anderson B. (1994) The Impact of Lobbying Laws and Their Enforcement: A Contrasting View. *Social Sciences Quarterly* 75(2): 378–381.

Holman C. and Luneburg W. (2012) Lobbying and Transparency: A Comparative Analysis of Regulatory Reform. *Interest Groups & Advocacy* 1(1): 75–104.

Holyoke T. T. (2015) *The Ethical Lobbyist: Reforming Washington's Influence Industry*. Washington, DC: Georgetown University Press.

Hrebenar R. J., Nakainura A. and Nakamura A. (1998) Lobby Regulation in the Japanese Diet. *Parliamentary Affairs* 51(4): 551–552.

Jordan G. (1998) Towards Regulation in the UK: From General Good Sense "to Formalized Rules". *Parliamentary Affairs* 51(4): 524–525.

King G., Keohane R. O. and Verba S. (1994) *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

Krippendorff K. (2004) Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research* 30(3): 411–433.

Landis R. and Koch G. (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33(1): 159–174.

Lombard M., Snyder-Dutch J. and Bracken Campanella C. (2002) Content Analysis in Mass Communication Assessment and Reporting of Intercoder Reliability. *Human Communication Research* 28(4): 587–604.

Lowery D. and Grey V. (1997) How Some Rules Just Don't Matter: The Regulation of Lobbyists. *Public Choice* 91(2): 139–147.

Mahoney J. and Goertz G. (2006) A Tale of Two Cultures: Contrasting Quantitative and Qualitative Research. *Political Analysis* 14(3): 227–249.

Maggetti M. and Gilardi F. (2016) Problems (and Solutions) in the Measurement of Policy Diffusion Mechanisms. *Journal of Public Policy* 36(1): 87–107.

McMenamin I. (2004) Varieties of Capitalist Democracy: What Difference Does East-Central Europe Make? *Journal of Public Policy* 24(3): 259–274.

Mikhaylov S., Laver M. and Benoit K. (2012) Coder Reliability and Misclassification in the Human Coding of Party Manifestos. *Political Analysis* 20(1): 79–91.

Munck G. L. and Verkuilen J. (2002) Conceptualizing and Measuring Democracy. *Comparative Political Studies* 35(1): 5–34.

Newmark A. J. (2005) Measuring State Legislative Lobbying Regulation, 1990–2003. *State Politics & Policy Quarterly* 5(2): 182–191.

Organisation for Economic Co-Operation and Development (OECD) (2008) *Lobbyists, Government and Public Trust: Building a Legislative Framework for Enhancing Transparency and Accountability in Lobbying*. Paris, France: OECD Publisher.

Organisation for Economic Co-operation and Development (OECD) (2012) *Lobbyists, Government and Public Trust, Volume 2*. Paris, France: OECD Publisher.

Opheim C. (1991) Explaining the Differences in State Lobby Regulation. *The Western Political Quarterly* 44(2): 405–421.

Ozymy J. (2010) Assessing the Impact of Legislative Lobbying Regulations on Interest Group Influence in US State Legislatures. *State Politics & Policy Quarterly* 10(4): 397–420.

Ozymy J. (2013) Keepin' on the Sunny Side: Scandals, Organized Interests and the Passage of Legislative Lobbying Laws in the American States. *American Politics Research* 41(1): 3–23.

Paxton P. (2000) Women in the Measurement of Democracy: Problems and Operationalization. *Studies in Comparative International Development* 35(3): 92–111.

Rechtman R. E. and Larsen-Ledet J. P. (1998) Regulation of Lobbyists in Scandinavia: A Danish Perspective. *Parliamentary Affairs* 51(4): 579–586.

Rocco P. and Thurston C. (2014) From Metaphors to Measures: Observable Indicators of Gradual Institutional Change. *Journal of Public Policy* 34(1): 35–62.

Rogers M. Z. and Weller N. (2014) Income Taxation and the Validity of State Capacity Indicators. *Journal of Public Policy* 34(2): 183–206.

Ronit K. and Schneider V. (1998) The Strange Case of Regulating Lobbying in Germany. *Parliamentary Affairs* 51(4): 559–567.

Rooduijn M. and Pauwels T. (2011) Measuring Populism: Comparing Two Methods of Content Analysis. *West European Politics* 34(6): 1272–1283.

Rosenson B. A. (2003) Against Their Apparent Self-Interest: The Authorization of Independent State Legislative Ethics Commissions, 1973–96. *State Politics & Policy Quarterly* 3(1): 42–65.

Rush M. (1998) The Canadian Experience: The Lobbyists Registration Act. *Parliamentary Affairs* 51(4): 516–523.

Seawright J. and Collier D. (2014) Rival Strategies of Validation: Tools for Evaluating Measures of Democracy. *Comparative Political Studies* 47(1): 111–138.

Stone D. (2004) Transfer Agents and Global Networks in the Transnationalization of Policy. *Journal of European Public Policy* 11(3): 545–566.

Thomas C. S. (1998) Interest Group Regulation Across the United States: Rationale, Development and Consequences. *Parliamentary Affairs* 58(4): 500–515.

True J. and Mintrom M. (2001) Transnational Networks and Policy Diffusion: The Case of Gender Mainstreaming. *International Studies Quarterly* 45(1): 27–57.

Veksler A. (2015) Diluted Regulations: A Need to Review the Theoretical Classification of the Different Lobbying Regulatory Environments. *Journal of Public Affairs* 15(1): 56–64.

Wagenaar A. C., Harwood E. M., Silianoff C. and Toomey T. L. (2005) Measuring Public Policy: The Case of Beer Keg Registration Laws. *Evaluation and Program Planning* 28(4): 359–367.

Warhurst J. (1998) Locating the Target: Regulating Lobbying in Australia. *Parliamentary Affairs* 58(4): 538–550.

Witko C. (2007) Explaining Increases in the Stringency of State Campaign Finance Regulation, 1993–2002. *State Politics & Policy Quarterly* 7(4): 369–393.

Woll C. (2012) The Brash and the Soft-Spoken: Lobbying Styles in a Transatlantic Comparison. *Interest Groups & Advocacy* 1(2): 193–214.

Yishai Y. (1998) Regulation of Interest Groups in Israel. *Parliamentary Affairs* 51(4): 568–578.