Pontifícia Universidade Católica do Rio de Janeiro

Clara Monteiro Vieira

A Pilot Study using Anchoring Vignettes and the Rasch Probabilistic Model: a contribution to democracy measurements

Dissertação de Mestrado

Dissertation presented to the Programa de Pós-graduação em Metrologia of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Metrologia.

Advisor: Prof. Elisabeth Costa Monteiro

Rio de Janeiro
October 2024

Pontifícia Universidade Católica do Rio de Janeiro

Clara Monteiro Vieira

A Pilot Study using Anchoring Vignettes and the Rasch Probabilistic Model: a contribution to democracy measurements

Dissertation presented to the Programa de Pós-graduação em Metrologia of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Metrologia. Approved by the Examination Committee:

Prof. Elisabeth Costa Monteiro
Advisor
Programa de Pós-graduação em Metrologia – PUC-Rio

Prof. William Paul Fisher Jr
University of California, Berkeley

Prof. Maurício Nogueira Frota
Programa de Pós-graduação em Metrologia – PUC-Rio

Prof. Carlos Roberto Hall Barbosa
Programa de Pós-graduação em Metrologia – PUC-Rio

Rio de Janeiro, October 4th, 2024

Clara Monteiro Vieira

Graduated in Social Sciences (Bachelor’s and Licentiate degrees) from the Pontifical Catholic University of Rio de Janeiro in 2021, with Additional Domain in ‘Evaluation and Management of Public Policies’.

Bibliographic data

Vieira, Clara Monteiro

A Pilot Study using Anchoring Vignettes and the Rasch Probabilistic Model: a contribution to democracy measurements / Clara Monteiro Vieira; advisor: Elisabeth Costa Monteiro. – Rio de Janeiro: PUC-Rio, Programa de Pós-Graduação em Metrologia, 2024.

93 f.: il. ; 29,7 cm

1. Dissertação (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Programa de Pós-graduação em Metrologia.

Inclui referências bibliográficas.

1. Metrologia – Dissertações. 2. Medições em Ciências Sociais. 3. Teoria de Medição de Rasch. 4. Medição de Democracia. 5. Incerteza de Medição. 6. Vinhetas de Ancoragem. I. Monteiro, Elisabeth Costa. II. Pontifícia Universidade Católica do Rio de Janeiro. Programa de Pós-Graduação em Metrologia. III. Título.

CCD: 389.1

To my friends and family

Acknowledgements

I would like to express my deep gratitude to Elisabeth for all her care, support, and patience, both as an advisor and as a mother. I also thank her for all the knowledge and values shared with me over the years. I also appreciate her courage in exploring the emerging field of Metrology applied to the Human and Social Sciences, venturing beyond her “comfort zone” (if she has such a thing) in Natural Sciences. Her interdisciplinarity and constant eagerness to learn are truly inspiring.

I am deeply honoured and grateful to William Fisher for his numerous and invaluable contributions, feedback, and guidance, which have also been particularly motivating. I am likewise extremely grateful to all the members of the examining board –namely, William Fisher, Mauricio Frota and Carlos Hall– for their kind and thoughtful feedback. I am sincerely honoured and grateful to Mauricio Frota and Carlos Hall for agreeing to review this work, despite the challenge of engaging with a research area (regarding psychosocial measurements) that differs from their usual focus.

I extend my gratitude to all those whose interest in this research project were especially motivating. I am also deeply honoured and grateful to Ricardo Ismael for his kind remarks and thoughtful contributions as member of the “examining board” at the presentation of the research topic of my master's dissertation.

I am truly honoured and grateful to the members of the V-Dem Project's Steering Committee for their expressed interest in this research and for kindly providing the requested database of responses to anchoring vignettes, which was examined in this study.

I am also profoundly thankful to Marcia, Paula, and Felipe, for their guidance and support.

I am especially grateful to my family, who have always stood by my side in the most joyful and humorous manner. I would also like to convey my profound appreciation to my dearest friends, from whom I continue to learn a great deal, even though we often do not stay much in touch.

This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil* (CAPES) - Finance Code 001.

The study was also financed in part by FAPERJ (*Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro*) –‘*Programa Bolsa Mestrado e Doutorado Nota 10*’.

Abstract

Vieira, Clara Monteiro; Monteiro, Elisabeth Costa (Advisor). **A Pilot Study using Anchoring Vignettes and the Rasch Probabilistic Model: a contribution to democracy measurements**. Rio de Janeiro, 2024. 93p. Dissertação de Mestrado - Departamento de Metrologia, Pontifícia Universidade Católica do Rio de Janeiro.

Measuring psychosocial phenomena, such as democracy, is challenging due to the influence of subjective perceptions and difficult-to-assess variables. It is crucial for democracy assessments to produce reliable and comparable results, as they play a vital role in analysing political realities and influencing decision-making at national and international levels. While essential elements of the Measurement Science are manifest throughout the discussions on democracy measurement, important aspects must be addressed to align the approaches used in this field with fundamental metrology precepts. Recent studies have suggested incorporating the psychometric approach known as the ‘Rasch model’ into the metrological system associated with each attribute as a potential solution to the challenges of providing comparable and reliable measures of psychosocial traits. However, this approach has yet to be explored in democracy measurement. This study applies the Rasch model to the democracy measuring system from the Varieties of Democracy project (V-Dem) and their recently incorporated anchoring vignettes database. It analyses coders' responses to the indicators constituting the V-Dem survey on Deliberation. The proposed method evaluates the system's performance, considering the coders as the sensing element and the indicators as the measuring system's structure. Sources of measurement uncertainty are discussed within the measurement model. The study reveals items associated with critical demands for revision based on multiple parameters, including differential item functioning based on coders' continent of origin and secondary dimensions affecting the sensor. Addressing these elements can contribute to enhancing the reliability of democracy assessment and advancing political science research.

Keywords

Metrology; Measurement in Social Sciences; Rasch Measurement Theory; Democracy Measurement; Measurement Uncertainty; Human Sensor.

Resumo

Vieira, Clara Monteiro; Monteiro, Elisabeth Costa (Advisor). **Um estudo piloto usando vinhetas de ancoragem e o modelo probabilístico de Rasch: uma contribuição para medições de democracia**. Rio de Janeiro, 2024. 93p. Dissertação de Mestrado - Departamento de Metrologia, Pontifícia Universidade Católica do Rio de Janeiro.

A medição de fenômenos psicossociais, como a democracia, é um desafio, dada a influência de percepções subjetivas e variáveis difíceis de medir. É fundamental que avaliações de democracia tenham resultados confiáveis e comparáveis, já que exercem um papel vital na análise de realidades políticas e influenciam a tomada de decisões a nível nacional e internacional. Embora elementos centrais da Ciência da Medição permeiem as discussões sobre a medição da democracia, aspectos importantes devem ser abordados para alinhar os métodos usados nesse campo com preceitos fundamentais da metrologia. Estudos recentes sugerem a incorporação da abordagem psicométrica conhecida como o ‘modelo Rasch’ no sistema metrológico associado a cada atributo, como potencial solução aos desafios de prover medidas comparáveis e confiáveis de grandezas psicossociais. Contudo, esta abordagem ainda não foi explorada na medição da democracia. Este estudo aplica o modelo Rasch ao sistema de medição de democracia do projeto '*Varieties of Democracy*' (V-Dem) e à sua base de dados de vinhetas de ancoragem recentemente incorporada. Analisam-se as respostas dos codificadores aos indicadores que constituem o *survey* do V-Dem sobre Deliberação. O método proposto examina o desempenho do sistema, considerando os avaliadores como o elemento sensor e os indicadores como a estrutura do sistema de medição. As fontes de incerteza de medição são discutidas no modelo do sistema de medição. O estudo revela itens com questões críticas para revisão com base em múltiplos parâmetros, incluindo o funcionamento diferencial dos itens conforme o continente de origem dos avaliadores e dimensões secundárias a afetar o sensor. O tratamento desses elementos pode contribuir à maior confiabilidade da medição de democracia e ao avanço da pesquisa em ciência política.

Palavras-chave

Metrologia; Medição em Ciências Sociais; Teoria de Medição de Rasch; Medição de Democracia; Incerteza de Medição; Sensor Humano.

Table of Contents

1 Introduction 13

1.1 Objectives 15

1.2 Master’s Dissertation Structure 16

2 Metrology in the Human and Social Sciences 18

2.1 Metrology in the early days of Social Sciences 20

2.2 Current endeavours for incorporating Metrology into Psychosocial Measurements 24

3 The Measurement of Democracy 29

3.1 Democracy measuring systems 30

3.2 Measurement Science and the Validity of Democracy Measurement 34

4 Metrological characterisation of the sensing elements of the V-Dem measuring system 37

4.1 Democracy measuring system: the "Varieties of Democracy" project (V-Dem) 37

4.2 Development of an approach for the metrological evaluation of the V-Dem's sensing elements 42

4.3 Modelling of the measuring system 44

4.4 Results from the proposed approach 47

5 Discussion, Conclusions and Future Work 65

5.1 Conclusions and Future work 77

References 79

List of figures

Figure 2.1. Diagram of correlating connections between macro-level variables to analyse causality associations with suicide rates in diverse contexts, within the Durkheim study [90] 22

Figure 4.1. Diagrams of the measuring instrument models associated: in (a), with the conventional application of the V-Dem in measuring the quality of democracy in a country; in (b), with V-Dem's use of anchoring vignettes as quality-level references; and, in (c), with the proposed model focused on assessing the coders' ability to identify the degree of a country's democracy attribute using anchoring vignettes. 45

Figure 4.2. Difficulty measures for the fourteen indicator-vignette combinations across the three samples of respondents analysed (MAX-MIN, MAX, MIN). Indicator-vignette combinations are labelled according to Table 4.2, and arranged from left to right in increasing order of difficulty. 48

Figure 4.3. Raters' and items' positions according to their Rasch measure and standardised infit value [4] for the MAX-MIN sample. The semi-transparent black circles represent individual coders, with darker regions indicating a more significant number of coders plotted in the same area. Pink circles display minimum-level vignettes, while blue circles indicate maximum-level vignettes. The diameter of each circle indicated along the vertical axis (in logit) denotes the random error associated with the corresponding Rasch measure. 52

Figure 4.4. Raters' and items' positions according to their Rasch measures and standardised infit values [4] for the MIN sample. Pink circles indicate minimum-level indicator-vignettes. See Figure 4.3 caption for a complete description of the illustrated elements. 53

Figure 4.5. Raters' and items' positions according to their Rasch measures and standardised infit values [4] for the MAX sample. Blue circles indicate maximum-level indicator-vignettes. See Figure 4.3 caption for a complete description of the illustrated elements. 54

Figure 4.6. Item difficulty measures estimated based on the three largest subsamples of coders by continent (A, C and D) in the MAX-MIN sample, compared with the item measures based on the full MAX-MIN sample (*), shown in Figure 4.2. 56

Figure 4.7. Item difficulty measures estimated based on the three largest subsamples of coders by continent (A, C and D) in the MIN sample, compared with the item measures based on the full MIN sample (*), shown in Figure 4.2. 56

Figure 4.8. Item difficulty measures estimated based on the three largest subsamples of coders by continent (A, C and D) in the MAX sample, compared with the item measures based on the full MAX sample (*), shown in Figure 4.2. 57

Figure 4.9. Items from the MAX sample ranked according to their difficulty measures (horizontal axis) and their residual loadings in the first PCAR contrast (vertical axis). On the right side of the graph, the items are grouped in three clusters according to their residual loadings in the first contrast. 59

Figure 4.10. Items from the MIN sample ranked according to their difficulty measures (horizontal axis) and their residual loadings in the first PCAR contrast (vertical axis). On the right side of the graph, the items are grouped in three clusters according to their residual loadings in the first contrast. 60

Figure 4.11. Items from the MAX-MIN sample ranked according to their difficulty measures (horizontal axis) and their residual loadings in the first PCAR contrast

(vertical axis). On the right side of the graph, the items are grouped in three clusters according to their residual loadings in the first contrast. 60

Figure 5.1. Distribution of raters as to their ability measures, in logit (horizontal axis), for each group of raters distinguished by their continents of origin (vertical axis), in the MAX-MIN sample; and the approximate position of item 'engage 1' (indicated by the dashed line) in that same scale (horizontal axis), according to its difficulty measure in the MAX-MIN sample. 68

Figure 5.2. Joint display of the Expected and Empirical Item Characteristic Curves (ICC) for 'engage 1' in the MAX-MIN sample, indicating the probability of correct responses (vertical axis) given the difference of rater ability relative to the item's difficulty (horizontal axis), based on the Rasch model –the Expected ICC (represented by the red curve)– or on the observed data –the Empirical ICC (the blue line connecting 'x' markers). The boundary lines around the Expected ICC show a (vertically interpreted) 95 % confidence band for the expected observations [162]. 69

Figure 5.3. Number of coders' responses assigning the indicator-vignette pairs 'reason 1' and 'reason 3' to each rating category of the corresponding indicator, in each sample of respondents (MAX-MIN, MIN, and MAX) 75

Figure 5.4. Number of coders' responses assigning the indicator-vignette pairs 'engage 1' and 'engage 5' to each rating category of the corresponding indicator, in each sample of respondents (MAX-MIN, MIN, and MAX) 76

Figure 5.5. Number of coders' responses assigning the indicator-vignette pairs 'countr 1' and 'countr 5' to each rating category of the corresponding indicator, in each sample of respondents (MAX-MIN, MIN, and MAX) 77

List of tables

Table 3.1. Democracy indices, their abbreviation, and related references 30
Table 3.2. Characterisation of democracy measuring systems as to their approach to the concept of democracy 31
Table 3.3. Characterisation of democracy measuring systems as to their aggregation rule and results' format 33
Table 4.1. Principles measured by V-Dem's Democracy Indices [24] 38
Table 4.2. Constituent indicators of the V-Dem survey on Deliberation, with their symbols and corresponding questions [24] 40
Table 4.3. Number of responding raters distributed according to their continent of origin and the coders' rating sample associated with the vignettes set at the lowest, the highest or both extreme construct levels, respectively indicated by MIN, MAX and MAX-MIN. 41
Table 4.4. Names used for each indicator-vignette combination, including the five indicators from the V-Dem index of the Deliberative Component of democracy (Db) and two indicators from the Egalitarian Component index (Eg), for the vignettes depicting the lowest level of each indicator's measured construct (Minimum), numbered 1, and those depicting the highest level of construct (Maximum), which are numbered 3 to 5 depending on the indicator's highest response category level. 42
Table 4.5. Probability of 'correct' rating to each indicator-vignette pair according to coder ability. 49
Table 4.6. Item statistics for the MAX-MIN sample, indicating the sum of correct responses (total score) for each of the fourteen items of the MAX-MIN sample, their Rasch measure of difficulty, random error (Model S.E.) and fit statistics –mean-square and standardised values of infit and outfit. The items (indicator-vignette combinations) are labelled according to Table 4.2. 51
Table 4.7. Rasch model reliability and separation indices for person and item measures, calculated from each sample of responses (MIN, MAX, MAX MIN) 54
Table 4.8. The variance explained by measures and the unexplained variance in the first PCA contrast of residuals, for each of the three main data samples used in this study (MAX-MIN, MAX, and MIN) 58
Table 4.9. Items grouped by clusters according to their residual loadings in the first PCAR contrast, for each of the three main samples (MAX, MIN and MAX-MIN) –as shown in Figure 4.9, Figure 4.10 and Figure 4.11. 61
Table 4.10. correlation of rater ability measures across the three clusters of items grouped by their residual loadings in the 1st PCAR contrast, for each of the three main data samples (MIN, MAX, and MAX-MIN) 62
Table 5.1. Number of responses associating each of the fourteen indicator-vignette pairs to each rating category of the corresponding indicator, in each sample of respondents (MAX-MIN, MIN and MAX). 72

"Ideas come when we do not expect them, and not when we are brooding and searching at our desks. Yet ideas would certainly not come to mind had we not brooded at our desks and searched for answers with passionate devotion."

Max Weber

1
Introduction

Fundamental precepts of Metrology apply to measurement practices across all fields of knowledge. An essential requirement is that of metrological traceability, which ensures the achievement of comparable measurement results for the same property, as each result “can be related to a reference through a documented unbroken chain of calibrations, each contributing to the measurement uncertainty” (JCGM 200:2012, section 2.41) [1]. While providing metrological traceability is naturally challenging, it becomes particularly more complex in the Human and Social Sciences, in which the intricate features of their measurands and measurement processes pose unique difficulties [2]-[15].

Subjective perceptions play a critical role in measuring psychosocial phenomena, as they can be the object of measurement or even a part of the measurement system [2], [5], [9], [10], [15]. The complexity of the measurement process is enhanced by the numerous variables that may affect it, as they are particularly hard to assess – an issue that is shared, to some extent, with other domains, such as chemistry, biology, and quantum physics [7], [16].

Likewise, the measured properties within the Human and Social Sciences are usually characterised by significant definitional uncertainty. Many concepts addressed in that field of science present highly abstract definitions [2], [3], [1], [15], which may also prove to be quite unstable, as they are primarily influenced by particular socio-historical contexts in which the concepts are used [2], [8]. Given their very abstract nature, such concepts’ definitions are also typically difficult to operationalise; thus, even when a certain definitional consensus is achieved, it often remains unclear how to empirically assess (and, therefore, measure) the concept at hand [3], [1], [15], [17].

Democracy measurement deals with those challenges. Despite many fluctuations in the way democracy has been defined throughout millennia [18], at the core of the concept, there is substantial agreement over its general meaning as “rule by the people” [1], [19]-[21]. Still, the definition of democracy in more specific terms remains the subject of much debate, and consequently, so does its measurement [1], [15], [20]-[23].

Measuring democracy may offer valuable insights for both academic and decision-making activities. Moreover, given its effect on public opinion, empirical assessments stipulating variations on the level of democracy across space and time may significantly impact local, national, and international political relations. Thus, providing reliable and comparable measures of democracy is of utmost importance. For that purpose, fundamental metrological directions applicable to measurement in all fields of scientific knowledge should be considered [7], [13]-[16].

Since the second half of the twentieth century, multiple approaches have been proposed to measure the democratic quality of a given country at a given time [1], [15], [19]-[35]. The measuring system of the "Varieties of Democracy" project (V-Dem), which had its first public data release in 2016 [24], stands out for its vast coverage and availability of data, as well as its extensive work on theoretical and methodological aspects [20], [36].

For much of its data, V-Dem relies on country experts to code several difficult-to-observe variables for characterising the actual state of affairs about the democratic features of a country's political regime over time [20], [21]. Each expert codes independently, revealing possible patterns of disagreement that V-Dem accounts for by applying Bayesian Item Response Theory (IRT) modelling techniques [21], [37]. However, each expert is able to rate no more than a few countries, limiting the available data for cross-country comparisons [21]. To overcome this issue, the V-Dem project, following the approach proposed by [38], recently incorporated 'anchoring vignettes', which are brief descriptions of hypothetical cases for coders to rate irrespective of their country of expertise, aimed at standardising levels for the property measured by each indicator [21].

Despite numerous efforts undertaken to evaluate and improve democracy measurement practices [1], [15], [21], [22], [28], [36]-[37], [39]-[43], fundamental metrological principles have yet to be fully integrated into existing approaches [15]. These approaches cannot provide a standard scale, based on a constant unit of reference, to compare the results of democracy measurement across countries, periods, or using different measuring systems [15]. Consequently, the objectivity of measurement results is undermined, in that the connection between the measurement results and the property under measurement is not clear, and intersubjectivity is also affected, in that there is no uniform interpretation of measurement results in different measurement contexts [8], [13]-[15], [44], [45]. This issue stems from the intricate nature of measuring psychosocial phenomena.

A metrological system linking measurement standards through traceability chains is required to ensure reliable and comparable measurement results of a given trait. As a potential solution to the challenges of providing comparable and reliable measures of psychosocial traits, recent studies have suggested incorporating the psychometric approach known as the Rasch model into a metrological system associated with each attribute [5], [6], [8], [10], [12]-[15], [46]-[49]. This would leverage the Rasch model's features to examine the objectivity and intersubjectivity of measurement results [8], [13], [45], enabling their comparison across different measurement contexts. However, the Rasch model remains unexplored in the literature on democracy measurement [15].

1.1 Objectives

This work aims to contribute to implementing the Rasch measurement theory into the framework for the metrological characterisation of democracy measuring systems. Focusing on the measuring system provided by the 'Varieties of Democracy' (V-Dem) project, a pilot study is set up to examine the performance of V-Dem expert coders –representing the sensing elements of the democracy property transducer– to detect the level of the construct associated with a given anchoring vignette –serving as the measurement standard for a level of each democracy-related property. The metrological characterisation of the detecting system is carried out by applying the Rasch model to a set of indicators from the V-Dem anchoring vignette database. In this preliminary investigation, the analysis is centred on vignettes anchored at the two extreme levels (minimum and maximum) of each indicator's measured construct.

The study aims to achieve the following specific objectives:

1. To review the literature on the foundations of Measurement Science and Measurement in the Humanities and Social Sciences.
2. To examine the literature on the conceptualisation and measurement of democracy, including existing democracy measuring systems and the strategies conventionally used to assess the quality of their results.

3. To investigate the existence of essential metrological elements in the democracy measurement framework and highlight potential advances and shortcomings in providing reliable and comparable democracy measurement results.
4. To survey the literature on applying the Rasch model to measure psychosocial properties and assess the quality of measurement results.
5. To select a democracy measuring system from among the most widely used (as identified in the literature), and with a database accessible upon request, to be utilised in this work.
6. To develop a model of the selected democracy measuring system and the strategy for its metrological characterisation, emphasising significant sources of measurement uncertainty.
7. To apply the developed strategy for metrological characterisation to a selected sample of the measuring system's data, focusing on experts' responses to the anchoring vignettes.
8. To analyse the suitability of the indicators used in the selected measuring system based on the analysis of testing parameters of the Rasch method, exploring the impacts of the information provided on knowledge of the measured phenomenon.
9. To discuss the possible limitations and contributions identified and provide suggestions for improving the measuring instrument to ensure metrological rigour and enable comparability of results obtained by different democracy measuring systems.

1.2 Master’s Dissertation Structure

This master's dissertation comprises six chapters, including the literature review, the theoretical basis, and the research methods in the first chapters, while the final chapters present and discuss the results of the developed study and summarise the conclusions.

Chapter 1 introduces the research topic, its significance, and the objectives of the present study.

Chapter 2 delves into the literature on Metrology applied to psychosocial measurements, and investigates possible metrological aspects addressed in the early days of Social Sciences. It outlines the currently proposed framework for promoting Metrology in psychosocial measurements, with a particular emphasis on the potential of the Rasch Measurement Theory for supporting metrological traceability in this complex field of application.

Chapter 3 reviews the existing democracy measuring systems, and the strategies conventionally used to assess the validity of democracy measurement results, identifying possible advances or yet-to-be-tackled features for incorporating metrology principles in democracy measurements.

Chapter 4 outlines the strategy developed for the metrological characterisation of the selected democracy measuring system –the V-Dem project– employing 'anchoring vignettes' as measurement standards and the Rasch probabilistic model. It provides a detailed description of the database, the method used, and the results obtained by applying the developed approach.

Chapter 5 presents the final considerations of the research, discussing points of convergence and divergence in the outcomes and previous studies in the literature. It concludes with the main contributions, suggestions, and plans for future research.

2
Metrology in the Human and Social Sciences

For promoting the quality of measurement aimed at properties across all fields of knowledge, recent studies in Measurement Science suggest object-relatedness (objectivity) and subject-independence (intersubjectivity) as two fundamental ideals that ought to be pursued [8], [13]-[15]. Object-relatedness comprises the extent to which the information obtained through measurement reflects only variations on the measured property. Enhancing objectivity thus requires a solid theory about the measurand and a reduced influence from other phenomena, minimizing definitional and instrumental uncertainties. Intersubjectivity, on the other hand, strives for consistent interpretation of measurement results by individuals across various locations and times, which depends on the metrological traceability of measurement results to the same reference scale [8], [13]-[15].

Providing metrological traceability of measurement results to the International System of Units (SI) is essential to ensure reliable and comparable quantity values in applications associated with all fields of knowledge. This aspect, however, has been a historical struggle since the early days, when efforts were directed to the elaboration of a metrological framework traditionally focused on promoting advances in the evolution of standards for measuring physical quantities.

After the signing of the '*Convention du Mètre*' (1875), the 1st '*Conference General de Poids et Mesures*' (CGPM), which took place in 1889, established international prototypes for physical quantities of length and mass units, respectively, meter and kilogram, also incorporating the second as the unit of time, according to astronomers' definition [50].

The high complexity of chemical and biological measurements, which also involve quantities belonging to the field of Natural Sciences, only much more recently received better attention and contributions to meet their metrological infrastructure demands [16], [51]-[54].

Metrological authorities' first initiatives toward meeting demands for chemical measurements took place with the adoption, in 1971, of the unit mole (symbol mol), for the quantity amount of substance, at the 14th CGPM, and the creation of the '*Comité consultatif pour la quantité de matière*' (CCQM), in 1993 [50].

In turn, measurements of biological quantities, which are particularly associated with even more challenging metrological demands, were addressed only at the 20th CGPM (1999) [16], [51]-[53]. However, unlike what happened in the case of chemical quantities, the metrological demands associated with biomeasurements did not receive specific support by creating a particular consultative committee for the area [16]. The responsibility for advancing the reliability of biomeasurements was absorbed by the CCQM, whose name was changed in 2014 to 'Consultative Committee for Amount of Substance: metrology in chemistry and biology' [16], [52].

Equally required and even more challenging is the global metrological framework to provide trustworthiness and comparability for measurements in Human and Social Sciences. Nevertheless, this issue has not yet been addressed in CGPM resolutions [14].

The sophistication of measurands associated with more complex areas involving Chemical, Biological, Human and Social Measurements requires dealing with the development of certified reference materials, creation of arbitrary units, and other alternative strategies to step forward to a metrological structure capable of harmonizing "nonphysical" measurements in all aspects of daily life demands [14], [16], [52].

Particularly regarding Human and Social Sciences, the influence of the subjective perceptions of researchers and research participants on the measurement process [5], [14], [15], [55] and difficulties in defining concepts [10], [14], [15], [17] are some of the elements of the complexity in the study of social phenomena. Such intricacies hinder but do not prevent initiatives to ensure reliability and comparability of measurement results in the Social Sciences and Humanities.

Recent studies have been endeavouring to meet the challenges associated with the complex characteristics of this scientific field [3]-[17], [37]-[49], [52], [56]-[85]. Among the current academic initiatives, it is worth mentioning the successful incorporation of measurements in Social Sciences among investigations addressed by the International Measurement Confederation (IMEKO) [13]-[15], [80]-[83], [85], being evidenced a massive effort of this scientific community to promote metrology in this field, including efforts to lead both physical and "nonphysical" measurement in a single, consistent concept system [8], [9], [13], [14], [52], [59], [80].

Despite the apparent novelty of the actions currently emerging to incorporate metrology concepts in Social Sciences, which aim to contribute to robust and comparable measurement results in this field of application, the literature indicates that the founding designers of sociology as an academic discipline had already expressed concerns regarding social measurements more than a century ago [14].

Section 2.1 discusses some of the first manifestations and concerns associated with essential elements of measurement science in the social sciences, and section 2.2 dives into the recent initiatives and efforts to effectively implement metrological requirements for the highly complex measurements in this field of scientific knowledge.

2.1 Metrology in the early days of Social Sciences

Emerging shortly after the intergovernmental metrological structure creation with the signing of the Metre Convention in 1875, the early methodological developments in the Social Sciences reflected ideas close to metrological concepts to ensure comparability as much as possible [14].

2.1.1 Emile Durkheim

The French sociologist Émile Durkheim (1858-1917, France) founded the first European department of sociology at the University of Bordeaux in 1895 [86]. Influenced by the positivist current of thought, Durkheim turned to the Natural Sciences – especially bioscience – when performing Social Science investigations [87], [88]. It is worth mentioning that both scientific fields share metrological challenges that still linger to the present time. With highly complex measurements, the measurement requirements framework in such fields of study is not yet adequately addressed or simply not at all. Interestingly, in his book from 1894 "*Les règles de la méthode sociologique*" [87], Durkheim already acknowledges such challenges that sociology has in common with biology, even though to a greater extent. As he states [87] (p.39): "*Tous ces problèmes qui, déjà en biologie, sont loin* *d'être clairement résolus, restent encore, pour le sociologue, enveloppés de mystère*”[1].

Building Natural Sciences’ analogies with the Social Sciences, Durkheim thought of society as an organism, whose parts (or “organs”) need to function well together to ensure the whole’s healthy functioning [87], [88]. Durkheim defined ‘social facts’ as his main object of study. ‘Social facts’ would be ways of feeling, acting, and thinking identifiable by three main traits such as generality, being applied to all members of a given society; exteriority from each individual, once they were not created by any particular person’s consciousness, but learned by people, generation after generation, and lasting much longer than the human lifespan; and coercivity, by which individuals are constrained into specific actions, not necessarily in conformity to each person’s intention [87]. Just as it is impossible to capture what is going on in someone’s mind by looking at each cell of their nervous system, Durkheim states that one wouldn’t be able to explain a social fact simply by looking at its manifestations in the individual level [87].

With a marked tendency toward an empirical approach, Durkheim used statistical strategies extensively. By increasing the number of cases whenever possible, the variable-oriented model of the comparative analysis performed by Durkheim aims to establish generalised connections between variables [89]. The general patterns pursuit guided Durkheim’s statistical approach to dealing with the time dimension from a transhistorical perspective [89]. Collective behaviors are, then, identified as an average effect of a variable by searching for statistical regularities of social facts [89]. Estimating the average effects of independent variables would allow investigating the ‘effects-of-causes’. Therefore, with the emphasis on generalizations over details, Durkheim establishes causality relationships, associating a phenomenon (social fact) to its cause or its effects (another social fact) [87].

For instance, in his famous study “*Le suicide: Étude de sociologie*” [90], performed with three religious’ communities (Protestants, Catholics, and Jews), Durkheim demonstrated that a social fact, the suicide rates, presented a statistical correlation with a macro-level variable constituted by the degrees of social integration, as illustrated in **Figure 2.1**. The statistical analysis allowed Durkheim,

[1] In English: “All these problems which, already in biology, are far from being clearly resolved, still remain, for the sociologist, shrouded in mystery”.

for example, to associate suicide rates to aspects of social context, whereas, contrary to what one might expect, there was no correlation with rates of psychopathology.

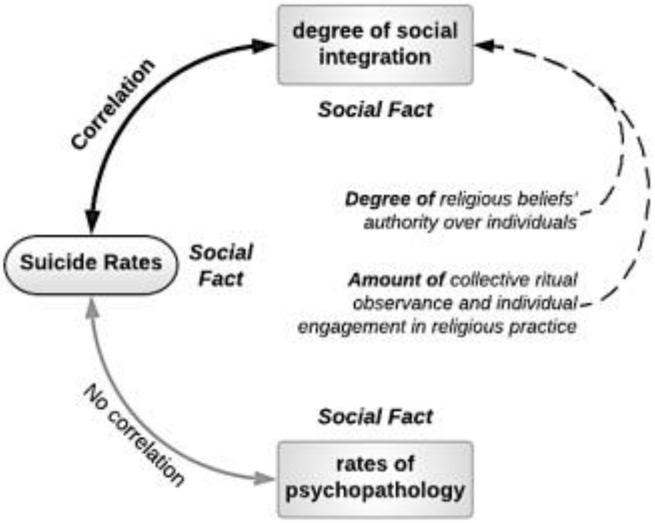
degree of social integration
Social Fact
Correlation
Degree of religious beliefs' authority over individuals
Suicide Rates
Social Fact
Amount of collective ritual observance and individual engagement in religious practice
No correlation
Social Fact
rates of psychopathology

Figure 2.1. Diagram of correlating connections between macro-level variables to analyse causality associations with suicide rates in diverse contexts, within the Durkheim study [90]

Looking for statistical regularities, Durkheim pursued stable objects as a condition for objectivity. In this sense, according to him, the more detached the "social facts" from their "individual" manifestation, the more objectively represented as a constant, minimizing subjective interference –as he states (p. 35) [87]:

> *"On peut poser en principe que les faits sociaux sont d'autant plus susceptibles d'être objectivement représentés qu'ils sont plus complètement dégagés des faits individuels qui les manifestent. En effet, une sensation est d'autant plus objective que l'objet auquel elle se rapporte a plus de fixité ; car la condition de toute objectivité, c'est l'existence d'un point de repère, constant et identique, auquel la représentation peut être rapportée et qui permet d'éliminer tout ce qu'elle a de variable, partant de subjectif"*[2].

[2] In English: "It can be stated as a principle that social facts are all the more likely to be objectively represented as they are more completely detached from the individual facts which manifest them. Indeed, a sensation is all the more objective as the object to which it relates has more fixity; for the condition of all objectivity is the existence of a point of reference, constant and identical, to which the representation can be related, and which allows to eliminate all that is variable, and therefore subjective, in it".

Durkheim’s quest for objectivity can be considered analogous to a pursuit towards minimizing the definitional and instrumental uncertainties of social measurements [14].

2.1.2 Gabriel Tarde

Gabriel Tarde (1843-1904, France), colleague of Durkheim and another influential figure in the foundation of the Social Sciences, also acknowledged the importance and complexity of quantification in this field of research, characterising this task as a new level of intellectual achievement [65]. Criticising Durkheim's statistical approach, which pursued general patterns, Tarde emphasised the significance of individual resonant effects in aggregate projections at the group level [65]. Recognising the intricacy of psychosocial measurements compared to those of the Natural Sciences, Tarde was concerned about the need for standardised units of measurement in the study of behaviour, cognition, and social relations [65].

In discussing the challenge of quantification in the study of psychosocial phenomena, Tarde elaborated on the distinction between qualitative (nominal) properties and quantities, with every quantity implying a similarity between opposed terms [91]. Describing all psychological states as combinations of “belief”, “desire” and “sensation” [92], Tarde also proposed that “belief” and “desire” were essentially quantities, but “sensation” was not, contributing to the inherent complexity in quantifying psychosocial phenomena [91].

2.1.3 Max Weber

Max Weber (1864–1920, Germany) introduced the first sociology department in his country, Germany, at the *Ludwig Maximilians Universität München*, in 1919 [93]. In contrast to Durkheim’s approach, Weber concentrated on grasping the intricacies of social phenomena from the micro-level, prioritizing subjectivity and meanings attributed to social actions [89], [94]-[97]. Weber's method, primarily qualitative in nature, sought to achieve an in-depth understanding of a complex unity through case-oriented comparisons, focusing on a small number of cases with a multitude of variables that interact within long-lasting processes [89].

As a tool for empirical analysis, the sociological approach proposed by Max Weber rested upon the development of the so-called Ideal Type, consisting of abstract concepts that provide a simplified representation of complex social realities, to be used as measurement standards for comparative analysis [89], [95], [98]. This strategy resembles the production of Reference Materials for chemical or biological measurements [14], considering the measurands for which the realization of SI units is still unavailable [16], [52], [54]. In these fields, it is possible to provide metrological traceability by developing Reference Materials with sufficient homogeneity and stability regarding specified properties, being established to be fit for their intended use in the measurement or examination of nominal properties [1]. Like the procedure using Certified Reference Materials as "primary reference standard," Weber's conception claims that the produced Ideal Types should be made available as a reference for further investigations of other cases –which would enable uniformity of interpretation through the *intersubjectivity* of measurement results [14].

Notably, Weber's concept of "ideal type" formed the basis for later developed measurement models addressing psychosocial properties [14], [68].

2.2 Current endeavours for incorporating Metrology into Psychosocial Measurements

When assessing the quality of measurement of psychosocial quantities, a commonly used concept is that of measurement validity, which states whether a measuring instrument, such as questionnaires or indicators, effectively measures the property it purports to measure [15], [41], [99], [100].

Some studies on measurement validity bring up concerns about the scale type in which a given data can be interpreted, as this limits the statistical techniques and mathematical operations that can be meaningfully applied to it [4], [15], [22], [41], [84]. Stevens, in 1946, proposed to distinguish between nominal, ordinal, interval, and ratio types of scales [84], each corresponding to different forms of representing and thus interpreting observations. In a nominal scale, observed events are classified into types with no inherent hierarchy among them, whereas an ordinal scale has categories arranged in order, indicating variations in a given underlying property [84]. Interval scales take a step further by preserving a consistent measurement unit across the scale and thus providing meaningful information on the differences between measurand values. Ratio scales additionally present absolute zeros that indicate the absence of the quantity being measured. Achieving measurement in a ratio scale is deemed likely beyond current capabilities for measuring most psychosocial phenomena [84]; therefore, efforts to enhance measurement in the Humanities and Social Sciences usually aim at providing information on the interval level. Nonetheless, the assumption that a particular dataset falls under an interval-level scale is often taken for granted [4].

A strategy for providing ordinal measurement based on nominal (qualitative) observations was proposed by Louis Guttman, in the 1940's [56]. Following the strategy based on 'ideal-types' proposed by the German sociologist Max Weber [2], Guttman developed a model embodying the conditions of "perfect" measurement [14], [68], posing unidimensionality as a central requisite –i.e. the condition according to which changes in the measurement indication reflect changes in a single quantity. Accordingly, in a Guttman scale, a series of observable attributes, typically assessed by dichotomous indicators, is hierarchically distributed according to variations in a single underlying quantity [1], [15], [41], [56]. This arrangement follows a cumulative pattern, meaning that an individual displaying any of these attributes is expected to display all lower-ranked (less "difficult") attributes as well.

The Guttman scale analysis operates under a deterministic measurement model, establishing a direct link between the property being measured (measurand) and the resulting measurement outcomes [6], [1], [15]. Deterministic approaches, however, are often deemed less suitable in complex measurement contexts, such as those found in the Human and Social Sciences, prompting researchers to opt for probabilistic strategies instead [6], [1], [15]. Rasch Measurement models and certain approaches in Item Response Theory (IRT) stand out as alternative tools in this scenario. By modelling the output of measurement as a probability distribution rather than a singleton, probabilistic approaches are useful for estimating meaningful differences along the measurement scale [6], [15]. Thereby, while Guttman scale analysis is deployed for producing ordinal measures; IRT and Rasch probabilistic models are used to attain measurements on the interval-level [6], [1], [15], [41].

Although the Item Response Theory and the Rasch measurement approach share similarities, they were developed independently from one another and have philosophical and methodological differences that are worth noting [101]. Primarily focused on describing idiosyncrasies of the data and explaining its variance, IRT methods used for providing interval-level data typically seek a model that best fits the data, incorporating one or more parameters designed to reflect characteristics of the sample [4], [10], [12], [15], [102]-[106]. In contrast, the Rasch Measurement approach tests the extent to which the data fits the model [4], [10], [12], [15], [102]-[106].

The Rasch Measurement Model provides a probabilistic realization of the Guttman scale. Hence, following an ideal-type conception of what would characterise measurement [14], [15], [68], it incorporates fundamental requirements of measurement into probabilistic frameworks. Besides 'unidimensionality', another requirement of the Rasch model is that of 'local independence', according to which responses to one item should not significantly influence responses to another [4], [15], [68], [85], [105]-[107].

With basic requirements built into the model, the Rasch measurement framework includes mechanisms to test the extent to which the data fit the model – a fundamentally different paradigm from that of similar approaches, such as Item Response Theory [4], [10], [12], [15], [85], [102]-[106].

The basic form of the Rasch probabilistic model, known as the dichotomous Rasch model, is based on two sets of parameters: one corresponding to each item (or indicator) in the measuring instrument and the other representing the individual instance of the quantity under measurement. Deriving from the Guttman scale, this particular model is designed for dichotomous data, where each item has two possible response options (e.g., correct/incorrect, present/absent), typically scored as 1 or 0.

In this model, as shown in equation (1), the probability that the individual *n* scores positively on item *i* –denoted as $\mathrm{P}(\mathrm{X_{ni}} = 1)$, or simply P_{ni}– is dependent on the difference between the individual's 'ability' (θ_n) –the measurand– and the item's 'difficulty' (δ_i) –the instrument's parameter [6].

$$P_{ni} = \mathrm{P}(\mathrm{X_{ni}} = 1|\theta_\mathrm{n}, \delta_\mathrm{i}) = \frac{e^{\theta_\mathrm{n}-\delta_\mathrm{i}}}{1 + e^{\theta_\mathrm{n}-\delta_\mathrm{i}}}. \tag{1}$$

In the Rasch model, the parameters' estimates are expressed in logits, which is the natural logarithm of the odds ratio derived from the raw probabilities of positive versus negative responses for each respondent or item [4], [6], [10], [68], [85], [105], [106], [108]. This relationship is formalised in equation (2).

$$\log_e\left(\frac{P_{ni}}{1-P_{ni}}\right) = \theta_\mathrm{n} - \delta_\mathrm{i}\,. \tag{2}$$

Hence, each logit represents a difference in the measurand value that increases the odds of observing the specified event by a factor of approximately 2.718 (the base of natural logarithms), with all logits maintaining uniform length relative to this change in odds [109]. In other words, this results in an interval-level scale, as defined by Stevens' classification [84], where a consistent unit of measurement is preserved, allowing for meaningful comparisons of differences between values across the scale.

In the Rasch Measurement Model, overall scores are considered sufficient statistics for estimating the parameters' values as long as the data fit the model [4], [10], [15]. Moreover, a unique feature of this approach is that it allows for parameter separability, meaning that item parameters can be estimated independently from the parameters related to the individuals under measurement and vice-versa [4], [6], [8], [10], [13], [15], [68], [79], [105], [106], [108], [110], [111]. This allows for measurement results to remain invariant, within a range of measurement uncertainty, across different measurement contexts, [4], [8], [13], [15], [68], [85]. This condition allows for intersubjectivity of measurement results [8], [15], as it enables comparisons of quantity values regardless of the individuals or stimuli that were instrumental for those comparisons. The attainment of invariant measurement on an interval scale is contingent upon the data meeting the measurement requirements set forth by the model –i.e., if the data sufficiently fit the model [4], [10], [15], [68], [79], [85], [105], [106], [108].

The family of Rasch measurement approaches includes other forms besides the dichotomous model depicted in equation (1), incorporating additional parameters, that regard, for instance, thresholds between response categories for polytomous response items (rating scale [112] or partial credit models [113]), or factors that are expected to influence measurement results in a systematic and measurable way (multifaceted model [114]). Even in those cases, the requirement of unidimensionality is maintained, since estimates for every parameter are placed along the same measurement scale –thus indicating magnitudes of the same underlying quantity– and this condition is empirically tested [4], [6], [10], [15], [68], [105], [111].

With a measurement model that is not contingent on the specific characteristics of the sample to which it is applied, Rasch analysis allows for psychosocial measurement to meet the same requirements applied to the measurement of physical quantities, demanding that measurement results related to the same measured property be comparable independently of the measurement object and the measuring instrument that is used. For that reason, the Rasch Measurement Theory has been appointed by recent studies in the field of Measurement Science as an ideal infrastructure for supporting metrological traceability in the Social Sciences [8], [9], [14], [15], [46]-[49], [59]. As described in [8], this could be structured by developing item banks aiming at building reference scales associated with each of the properties, in combination with Rasch model fitting [8], [13], [14].

3
The Measurement of Democracy

Initiatives to measure democracy have emerged since the latter half of the twentieth century [24]. With a history stretching back thousands of years, the word 'democracy' can be traced back to ancient Greece, from the combination of *demos* (people) and *kratos* (government). Despite the considerable variations in the way democracy has been conceived over time [18], its generic meaning of 'government by the people' has prevailed at the core of the concept [1], [19], [20], [24]. Its definition in more specific terms, however, remains the subject of much debate –as does its measurement [1], [20]-[23].

A number of systems have been developed to measure the democratic quality of a country at a given time [1], [15], [19]-[36], and several studies have also delved into the quality of these measurement approaches [1], [15], [24], [22], [24], [28], [36]-[37], [39]-[43]. However, challenges stemming from the inherent complexity of measuring psychosocial phenomena complicate these endeavours –which grapple with difficulties in defining concepts or assessing key variables, along with the significant impact of subjective perceptions on the measurement process [2]-[15].

Potentially providing valuable insights to describe, compare and explain political realities, democracy measurement is useful for informing research and decision-making processes. Moreover, given their impact on public opinion, empirical assessments stipulating variations on the level of democracy across space and time may significantly affect local, national, and international political relations. Thus, ensuring reliable and comparable results from democracy measurement is of utmost importance. For that purpose, fundamental metrological directions applicable to measurement in all fields of scientific knowledge should be considered [7], [13]-[16]. This chapter examines the current state of the art in developing and evaluating democracy measuring systems to identify potential gaps for incorporating metrology principles in this measurement field.

3.1 Democracy measuring systems

Some of the most prominent democracy indices currently available are briefly examined in this section. For each index, the name, abbreviation, and relevant references are listed in **Table 3.1**.

Table 3.1. Democracy indices, their abbreviation, and related references

Democracy Index	Index Abbreviation	Related References
Boix-Miller-Rosato	BMR	[31]
Democracy Barometer	DB	[27], [30], [33]
Democracy-Dictatorship	DD	[29], [32]
Freedom House status of freedom[3]	FH	[34]
Lexical Index of Electoral Democracy	LIED	[1], [19]
Polity 2 index	Polity 2	[35]
Unified Democracy Score	UDS	[28]
Vanhanen's index	Vanhanen	[26]
Varieties of Democracy indices[4]	V-Dem	[20], [21], [24], [115]-[117]

While not an exhaustive list of democracy measuring systems, **Table 3.1**'s selection of indices is sufficient to illustrate differences between their approaches. **Table 3.2** and **Table 3.3** elaborate on such differences, characterising each index according to a set of parameters [22], [23]. These parameters include, for **Table 3.2**, the breadth of elements addressed as observable features of democracy, which hinges on the way that concept is defined (*concept definition*), and the means used for data collection, which entails specific arrangements of the measuring system (*source of data*). In **Table 3.2**, democracy indices are characterised in terms of their *aggregation rule* –the approach used to combine the values of a set of indicators into a single value of an index– and the *presumed type of scale* of their main results.

[3] Despite focusing on "freedom" rather than democracy per se, the main index provided by the Freedom House organization (as part of their "Freedom in the World" annual reports) has been incorporated in this overview due to its widespread recognition as a measure of democracy.

[4] The V-Dem project offers five separate democracy indices (electoral, liberal, participatory, egalitarian, and deliberative), whereas the electoral democracy index is constitutive of all the other four indices.

Table 3.2. Characterisation of democracy measuring systems as to their approach to the concept of democracy

Index		concept definition			source of data				
		narrow	*mid.*	*broad*	*factual data*	*in-house coders*	*consulted expert*	*mass surveys*	*other indices*
BMR		[23]				x			
DB			[27] [30]					x	x
DD		[1] [20] [23] [29] [32]			x	x			
FH				[22] [23] [20]		x	x		
LIED		[1] [19]	[23]		x	x			
Polity 2		[20] [22] [118]		[23]		x			
UDS				[23]					x
Vanhanen		[22] [23]			x	x			
V-Dem	**electoral**		[23]	[32] [116]	x	x	x		
	liberal			[116]	x	x	x		
	participatory			[116]	x	x	x		
	deliberative			[116]			x		
	egalitarian			[116]			x		

The definition of democracy's key observable features, conditioning the choice of indicators, is a major point of debate [1], [20]-[23]. Previous studies on this matter distinguish between approaches based on 'minimalist' (*narrow*) and 'maximalist' (*broad*) definitions of the democracy concept. Accordingly, *narrow* definitions would encompass a limited range of variables as indicators of democratic quality, omitting features deemed important by others, while *broad* definitions would include a larger set of observable traits as defining features of democracy, at the risk of lacking empirical referents or compromising the analytical use of measurement results [22]. Some approaches have also been characterized as falling between 'maximalist' and 'minimalist' conceptions, being labelled '*mid.*' in **Table 3.2**.

However, assessments of the 'minimalist' or 'maximalist' nature of democracy measuring systems' conceptual approaches may often diverge [15], as they are contingent upon one's evaluation of what is theoretically relevant to the democracy concept. In this respect, the incompatible classifications attributed to the Polity 2 index, for instance, are notable, as its conceptual framework has been considered ‘minimalist’ by some [22], [118], and ‘maximalist’ by others [23] (**Table 3.2**). These conflicting interpretations illustrate the theoretical disagreements surrounding democracy’s defining features.

The specific arrangements of measuring instruments used by the different democracy measuring systems are also outlined in **Table 3.2** (*source of data*) [20]. Some of the indicators used in democracy measurement are primarily based on ***factual data***, such as the share of the population with the right to vote. Others present ratings that are more heavily influenced by value judgments, either by ***consulted experts*** or by members of the measurement project (‘***in-house coders***’) [20]. Also centred on subjective data, a less frequent approach is that of ***mass surveys*** designed to capture citizens’ behaviours and opinions that are theorised to reflect democratic or non-democratic aspects in their daily life experiences [20], [24], [30]. Furthermore, some democracy measuring systems, like those of UDS and DB, use data from ***other indices*** [20], [24], [28], [33].

Differences in the *aggregation rule* used to combine data from various indicators into a single index of democracy are illustrated in **Table 3.3**. Analogous to a measurement model, which is defined in the International Vocabulary of Metrology (VIM 3) as the “mathematical relation among all quantities known to be involved in a measurement” (JCGM 200:2012, section 2.48) [1], an aggregation function relies on a theoretical understanding of the measurand and its relation with other properties. The literature on the aggregation strategies applied to democracy measurement commonly distinguishes between two theoretical approaches underpinning different mathematical procedures [20], [24], [23], [36], [116], [117]. On the one hand, additive aggregation techniques, like sums, averages, and weighted averages, align with the view that indicators are mutually compensating and, therefore, partially substitutable in democracy measurement. This means that one aspect *or* another is included among the defining features of a democratic regime. On the other hand, if indicators are seen as interdependent or necessary conditions for democracy –requiring the presence of one attribute *and* another–, multiplicative aggregation procedures are justified[5]. Some democracy indices rely on both strategies: for instance, DB applies an additive approach at lower

[5] The methods used by BMR, DD and LIED for scoring democracy are identified, in **Table 3.3**, with a ‘multiplicative’ approach even though multiplication is not involved, due to the fact that indicators are treated as necessary conditions for achieving a given score [1], [29], [31].

aggregation levels and a multiplicative formula at higher levels [33], while V-Dem combines both techniques within a single function, obtaining their main democracy indices through an equally weighted average between the results from an additive and a multiplicative approach [21], [24]. Item Response Theory and Bayesian factor analysis are additionally employed in the construction of the V-Dem indices [21]. Likewise, UDS uses a Bayesian latent variable model to aggregate several democracy indices into a single index, assuming they all provide approximations of the same underlying quantity [28].

Table 3.3. Characterisation of democracy measuring systems as to their aggregation rule and results' format

Index	**aggregation rule**			**presumed type of scale**	
	ADDITIVE (*or*)	MULTIPLICATIVE (*and*)	OTHER METHODS	ORDINAL (*categorical*)	INTERVAL (*continuous*)
BMR		x		Binary	
DB	x	x			Free scaling criteria
DD		x		Binary	
FH	x			Three categories	
LIED		x		Seven categories	
Polity	x			Integers from -10 to 10	
UDS			Bayesian model		Standard normal distribution
Vanhanen		x			Continuous: 0-100
V-Dem 5 indices	x	x	IRT, Bayesian factor analysis		Continuous: 0-1

Table 3.3 also indicates the type of scale in which each measurement project explicitly or implicitly assumes to present its results. Some democracy measuring systems, like LIED and V-Dem, engage in notable discussions on the appropriate method for conveying information on specific types of scale [1], [19], [24]. The presumed scale type of the results from the Polity 2 index, in turn, remains ambiguous, as they are presented as a discrete distribution of integers ranging from -10 to +10 [35], but with no explicit indication on whether these numbers should be interpreted as categories on an ordinal scale or as equidistant values on an interval scale. Recent changes in the reporting methods of some democracy measurement

projects are also worth noting. The Democracy Barometer (DB) system has recently abandoned its fixed 0 to 100 scale, leaving it to researchers to decide the scaling criteria according to their specific purposes [33].

3.2 Measurement Science and the Validity of Democracy Measurement

Although lacking explicit reference to metrology as such, fundamental concerns with elements from the Science of Measurement are manifest throughout the discussions on democracy measurement [15]. Intimately related to the matter of subject-independence of measurement results (intersubjectivity), concerns about comparability frequently arise, with different democracy measuring systems being said to yield conflicting empirical findings [24], [29], [119]-[121]. Likewise, conceptual disagreements on the way democracy is defined have been outlined [20], [23], [36], [118], [122], and it has been emphasised the need for democracy measuring systems to rely on a definition of democracy whose applicability across space and time is explicitly justified [24].

Concerns about objectivity in democracy measurement are also evidenced as scholars draw attention to potential bias in the way democracy has been defined, which is likely to echo international asymmetries of power [24], [123]-[125]. In a similar vein, there is a growing call to move beyond strictly institutional definitions and embrace more inclusive and socially oriented conceptions of democracy [24], [122], [126]-[128]. Researchers have also questioned the feasibility of representing a particular operationalised concept of democracy as a singular variable, defined along a single (unidimensional) measurement scale [22], [43], [118], [129]-[133]. Discussions regarding the impacts of methodological choices over measurement error [22], [130], alongside attempts to investigate and minimise those errors [20], [23], [28], [39], [134]-[136], are likewise noticeable in the literature on democracy measurement.

Different strategies have been employed to assess the validity of democracy measurement results [1], [15], [28], [36], [39]-[43], [119], [120], [129], [134]-[148]. In investigating the approaches to measurement validity in Political Science research, [41] identifies four main traditions, each of which has applications in the field of democracy measurement: the 'case-based' tradition, the 'pragmatic' tradition, the 'Structural Equation Modelling with latent variables' (SEM-L), and the 'Levels of Measurement' tradition (LoM).

The case-based method, which can be predominantly qualitative, involves conducting in-depth case studies to gather more knowledge and evaluate whether scores accurately capture the realities they represent [41]. Studies that assess the validity of democracy measurement results through this approach include [138]-[145]. Valuable for understanding specific cases, this method can contribute to relevant conceptual and methodological insights. Nevertheless, this approach alone does not allow for systematic comparisons across various contexts.

The pragmatic tradition, in turn, uses straightforward statistical techniques, such as correlation and regression analysis, to evaluate the performance of indicators, based on immediate application purposes and regardless of any general measurement model [15], [41]. In the field of democracy measurement, methods aligned with this tradition have been used in [120], [137], [145]. Although fruitful for exploratory studies, the pragmatic tradition for assessing measurement validity has been criticised for its insufficient consideration of the connections between the measured property and the indication values [41]. Such a characteristic undermines object-relatedness and subject-independence of measurement results, as it limits the possibility of meaningful comparisons across different measurement contexts [15].

Structural Equation Modelling with Latent Variables (SEM-L), on the other hand, applies sophisticated statistical models for aggregating indicators and assessing measurement error [41]. SEM-L played a central role in the history of democracy measurement, as its extensive use in the 1980s, following the work of Bollen [148], has paved the way for the methodological sophistication of democracy measurement practice [25]. Studies applying Structural Equation Models for constructing and validating democracy measurement results include [100], [119], [134]-[136], [148].

Lastly, the Levels of Measurement tradition (LoM) brings up concerns about the scale type in which a given data can be reliably interpreted, as this limits the statistical techniques and mathematical operations that can be applied to the data [22], [41]. Studies developed under this tradition typically attempt to transform data for incorporating higher levels of measurement (scale types), thus broadening the range of applicable statistical techniques [41]. Methods used by this tradition include the Guttman scale analysis, the Rasch Measurement Theory, and the Item Response Theory (IRT).

In democracy measurement literature, Guttman scale analysis has been used to construct and validate ordinal measurement results [129], [144], [147]. Furthermore, a review study from 2002 on the democracy measurement literature underscored the relevance of the Guttman scale approach for empirically testing the unidimensionality hypothesis of a resultant scale [22].

For constructing interval-level measures of democracy, methods aligned with Item Response Theory have been applied [21], [37], [117], [136]. In contrast, the use of the Rasch model –which holds the potential for overcoming metrological challenges in the Human and Social Sciences– is practically absent in democracy measurement studies [15].

Nonetheless, an application of the Rasch model to assess measures of "political trust" from a widely used cross-national survey database that forms a specific part of the Democracy Barometer measuring system already demonstrates the promising features of this approach [149]. Despite not delving into metrological concepts, such as the potential of Rasch modelling to reach intersubjectivity by providing metrological traceability to measurement results, [149] evaluated unidimensionality, equivalence, and item hierarchy of political trust, revealing the lack of cross-national correspondence of political trust measurement results. These outcomes corroborated with predominant views on the theoretical literature indicating the non-unidimensionality of ‘political trust’, and contradicted the conventional political trust measurement practices, which typically assumed the unidimensionality of the construct. Hence, [149] points to the need for higher consistency and robustness in data analysis, and illustrates some promising features of the Rasch measurement approach, including its potential for investigating theoretical conceptions.

4
Metrological characterisation of the sensing elements of the V-Dem measuring system

In the context of a multitude of democracy measuring systems [1], [15], [19]-[36], the "Varieties of Democracy" project (V-Dem) was selected for this study's analysis. Released in 2016, V-Dem stands out due to its wide-ranging scope, database accessible upon request, and substantial research on theoretical and methodological topics [20], [36].

This chapter presents the study aimed at analysing the performance of the V-Dem transducer's sensing element, represented by the expert coders, in detecting the construct level of a given vignette, which serves as the measurement standard of a democracy property level. The metrological characterisation of the detecting system is performed by applying the Rasch model to a set of indicators from the V-Dem anchoring vignette database.

4.1 Democracy measuring system: the "Varieties of Democracy" project (V-Dem)

With a multidimensional approach to the democracy concept, V-Dem produces five main indices covering different democratic facets of a country's political regime: electoral, liberal, participatory, deliberative and egalitarian [20], [24]. The core values of these principles are presented in **Table 4.1** [24]. To assess these principles, V-Dem uses a large number of indicators, some of which are based on ordinal ratings given by consulted experts on difficult-to-observe variables [20], [24] –as previously highlighted in **Table 3.2**.

Table 4.1. Principles measured by V-Dem's Democracy Indices [24]

Principles	Description
Electoral	Based on Robert Dahl's concept of "polyarchy" [150], [151], the electoral principle of democracy focuses on making rulers accountable to citizens through **periodic elections**.
Liberal	The liberal principle of democracy focuses on protecting **individual and minority rights** against a "tyranny of the majority" and state oppression. It relies on constitutionally protected civil liberties, rule of law, and limits to the use of the executive power.
Participatory	The participatory principle of democracy emphasizes active citizen involvement in political processes, through elections and nonelectoral forms of **participation**.
Deliberative	The deliberative principle of democracy values decision-making informed by **respectful** and **reason-based** dialogue, prioritizing the **public good** over emotional or biased interests.
Egalitarian	The egalitarian principle of democracy posits that all groups should have **equal capabilities** to participate, serve, and influence policymaking, considering that inequalities in health, education, or income hinder the exercise of political rights.

A global network of around 4000 experts currently contributes to the V-Dem database [21]. The indicators they assess are distributed as questions along fifteen surveys (questionnaires), each of which tailored to a specific area of expertise [21], [37]. Each expert codes independently, and at least five are usually sought to rate a country on a given indicator, most of whom being nationals or residents of the country they rate [21]. This independent coding allows for possible patterns of disagreement to arise, which V-Dem accounts for by the use of Bayesian Item Response Theory (IRT) techniques [21], [37]. However, each expert provides ratings for only one or a few countries, limiting the data for cross-country comparisons [21]. Seeking to overcome these data constraints, V-Dem has recently adopted the use of "anchoring vignettes", brief descriptions of hypothetical cases that experts can rate irrespective of their country of expertise [21]. Analogous to measurement standards, these hypothetical cases are theorised to represent specific levels of the property measured by an indicator [38]. In the V-Dem project, vignettes were designed to correspond to idealised thresholds between adjacent categories of an indicator's rating scale, meaning the cases they illustrate should fit into either of these adjacent response options [152].

The present study aimed to analyse the performance of the V-Dem transducer's sensing element, represented by the expert coders, in detecting the construct level of a given vignette, which serves as the measurement standard of a democracy property level. Considering the evaluator as the measuring system's sensor and the synthetic reference texts ("anchoring vignettes") as reference materials for the property's level measured by each indicator, an approach was developed for the metrological characterisation of these sensors. This pilot study focused on raters' ability as the measurand for the proposed measuring instrument design rather than democracy levels.

Due to limitations in the size of the database supported by the software used in this analysis [4], experts' responses to a limited set of V-Dem indicators were considered. As this preliminary research focused on evaluating the performance of raters in classifying the descriptions associated with predefined construct levels – and not on measuring these construct levels themselves–, the indicators were selected to encompass the entire group of items from a particular V-Dem expert survey, even if they did not form a single index.

Of the V-Dem indices focusing on each of the five core principles of democracy distinguished by the project (**Table 4.1**), the deliberative is the only for which all indicators, also the fewest in number, are comprised in a single expert survey –the Deliberation survey (**Table 4.2**) [24]. Therefore, the present study focused on the complete set of indicators that constitute the V-Dem Deliberation survey, shown in **Table 4.2**. Most of its questions are either centred on the quality of discourse from political leaders or the general nature of public policy [24]. The survey comprises seven indicators, five of which form the V-Dem index of the Deliberative Component of democracy; while the remaining two, both related to aspects of public policy, are included in the index of democracy’s Egalitarian Component [24]. As shown in **Table 4.1**, V-Dem defines the deliberative quality of democracy as stemming from the ideal that policymaking at all levels is informed by respectful and reasoned dialogue aimed at the common good [20]. The egalitarian principle, in turn, would pertain to the distribution of power and resources, considering material and immaterial inequalities as significant hindrances to the *de facto* exercise of formal rights and freedoms [20].

Table 4.2. Constituent indicators of the V-Dem survey on Deliberation, with their symbols and corresponding questions [24][6]

Principle	Indicators (Deliberation survey)	Question
Deliberative	Reasoned justification v2dlreason	When important policy changes are being considered, i.e. before a decision has been made, to what extent do political elites give public and reasoned justifications for their positions?
	Common good v2dlcommon	When important policy changes are being considered, to what extent do political elites justify their positions in terms of the common good?
	Respect counterarguments v2dlcountr	When important policy changes are being considered, to what extent do political elites acknowledge and respect counterarguments?
	Range of consultation v2dlconslt	When important policy changes are being considered, how wide is the range of consultation at elite levels?
	Engaged society v2dlengage	When important policy changes are being considered, how wide and how independent are public deliberations?
Egalitarian	Particularistic or public goods v2dlencmps	Considering the profile of social and infrastructural spending in the national budget, how "particularistic" or "public goods" are most expenditures?
	Means-tested v. universalistic policy v2dlunivl	How many welfare programs are means-tested and how many benefit all (or virtually all) members of the polity?

From the set of anchoring vignettes representing the different levels of the construct assessed by each of the seven indicators examined in this study (**Table 4.2**), the two vignettes relating to the minimum and maximum magnitude of the categorical scale for each indicator were selected. The selection was motivated not only by the limited size of the database supported by the software but also by the possibility of assessing the adequacy of the coders' interpretation in classifying the level of the latent trait in the presence of significant differences in stimulus intensity, i.e. with high contrast in the magnitude of the rated property.

[6] For the complete text of the questionnaire items –including the rating options description– from the V-Dem Deliberation survey, see [24] (p. 169-172).

Three major data samples were considered: one with coders' responses to vignettes anchored at the minimum-level threshold of each indicator's categorical scale (MIN), another with the responses to vignettes at the maximum-level thresholds (MAX), and a third with joint responses to both sets of vignettes (MAX-MIN). Only coders with complete response strings for each sample were included. Therefore, the MIN sample covered the respondents who rated all minimum-level vignettes but not necessarily the maximum-level ones, while the MAX sample corresponded to the opposite condition. The number of coders and their distribution in the five continental regions of their country of specialization are shown, for each of the three samples, in **Table 4.3**. The five continents of origin of the evaluators are anonymised and specified by letters.

Table 4.3. Number of responding raters distributed according to their continent of origin and the coders' rating sample associated with the vignettes set at the lowest, the highest or both extreme construct levels, respectively indicated by MIN, MAX and MAX-MIN.

Continent	MIN	MAX	MAX-MIN	total
A	50	49	49	50
B	22	20	20	22
C	32	31	31	32
D	41	39	38	42
E	4	4	4	4
Total	149	143	142	150

In the present study, expert ratings of latent variables are considered the output of a measuring system consisting of different human sensors (the expert raters) interacting with a set of questionnaire items (the indicators). The system was redesigned to assess the measuring instrument performance in identifying the construct level represented by each anchoring vignette. Each vignette was thereby paired with its corresponding indicator as essential stimuli to which experts respond with their ratings.

The names used in this paper for each indicator-vignette combination are listed in **Table 4.4**. The titles of the indicators and their corresponding abbreviations were based on V-Dem’s codebook [24]; while the vignettes were numbered according to the threshold they represented on the indicator's rating scale. Vignettes set at the lowest threshold (minimum) were numbered 1, while those set at the highest threshold (maximum) were numbered 3 to 5, depending on the number of categories comprised in the indicator's response scale.

Table 4.4. Names used for each indicator-vignette combination, including the five indicators from the V-Dem index of the Deliberative Component of democracy (Db) and two indicators from the Egalitarian Component index (Eg), for the vignettes depicting the lowest level of each indicator's measured construct (Minimum), numbered 1, and those depicting the highest level of construct (Maximum), which are numbered 3 to 5 depending on the indicator's highest response category level.

	Indicators	Minimum	Maximum
Db	Reasoned justification	reason 1	reason 3
	Common good	common 1	common 4
	Respect counterarguments	countr 1	countr 5
	Range of consultation	conslt 1	conslt 5
	Engaged society	engage 1	engage 5
Eg	Particularistic or public goods	encmps 1	encmps 4
	Means-tested v. universalistic policy	univl 1	univl 5

4.2 Development of an approach for the metrological evaluation of the V-Dem's sensing elements

For the proposed approach, the polytomous ratings assigned by coders to each indicator-vignette pair were converted into dichotomous scores, indicating 'correct' or 'incorrect' responses according to the vignette's predefined construct level. Considering the V-Dem vignette design [152], correct responses were allowed to span two adjacent response options, thus including the two categories at the top of the indicator's ordinal scale for maximum level vignettes and the two at the bottom for minimum level vignettes. In this scenario, 'incorrect' responses to minimum level vignettes likely indicate that coders were more lenient in assessing the latent constructs compared to the perception that guided the vignette's design. Conversely, coders with 'incorrect' responses to maximum level vignettes would likely be more rigorous, as they assign lower construct levels to the hypothetical cases than the cases were intended to represent.

The resulting dichotomous scores were analysed using a dichotomous Rasch model –shown in equation (1), in chapter 2. The analysis was conducted using Bond&FoxSteps3 [4] –which is a smaller version of Winsteps [153], software produced by Michael Linacre, one of the Rasch models' main developers.

Following equation (1), the performance of the V-Dem sensor system, represented by the probability that experts' ratings of vignettes match the vignettes' predefined construct levels, was modelled in terms of the raters' ability to classify the vignettes as expected (θ_n) –the measurand– and the difficulty of each indicator-vignette pair in eliciting the intended responses (δ_i) –the instrument's parameter. Both ability and difficulty measures were produced along the same interval scale, relative to the degree of a presumed property with systematic influence over the probability of raters (the measurement sensors) identifying the construct level that each vignette was designed to span.

The distribution of raters' abilities and items' difficulties along this scale was examined. Rasch reliability (and separation) index, which refers to the proportion of variance associated with the construct being measured [4], [10], [68], [105], [106], was analysed for both item and person parameters. The software's indication of "Standard Error", representing levels of random error associated to each individual measure [10], is also taken into account. This specific indication hinges on the level of information an item (measuring instrument) reveals about the person parameter (examinee), or vice-versa, depending on their respective positions along the measurement scale [10].

The fit of the data to the Rasch measurement model was assessed through examination of the model residuals, which are the differences between the observed and expected performances of a given item or examinee [4], [68], [105], [106]. This study analysed the amount and likelihood of item residuals through slightly different forms of fit statistics, namely, the mean-square and standardised values for the infit and outfit statistics. In the infit indices, greater weight is given to the performance of individuals whose abilities are closer to the item's difficulty. This assertion is based on the premise that these individuals' performance should provide more sensitive information regarding their true abilities on that item [4], [68], [105], [106], [153]-[157].

To evaluate the unidimensionality of the data, a principal component analysis of item residuals was conducted, investigating if correlations among residuals are random or if they form a pattern suggesting an additional dimension that was not measured by the model [4], [10], [105], [106], [153]-[157].

The invariance of item difficulty was further assessed by testing for differential item functioning (DIF) across groups of coders from different continents. If the estimate of an item's difficulty varies more for each subsample than its error, this would indicate that the item does not function consistently across these groups of individuals [4], [68].

Furthermore, the study investigated the potential sources of uncertainty affecting the various stages of the measurement process in both the proposed and conventional designs of the measuring system.

4.3 Modelling of the measuring system

The diagram presented in **Figure 4.1** provides a simplified framework of the measuring system modelling [13], [158], considering the conventional approach employed by the V-Dem project to measure quantities associated with the quality of democracy (**Figure 4.1a** and **Figure 4.1b**), in parallel with the model proposed here to assess its sensing device (evaluators' ability) as the measurand (**Figure 4.1c**). The conceivable sources of measurement uncertainty and their location in the measurement process are also illustrated in **Figure 4.1**.

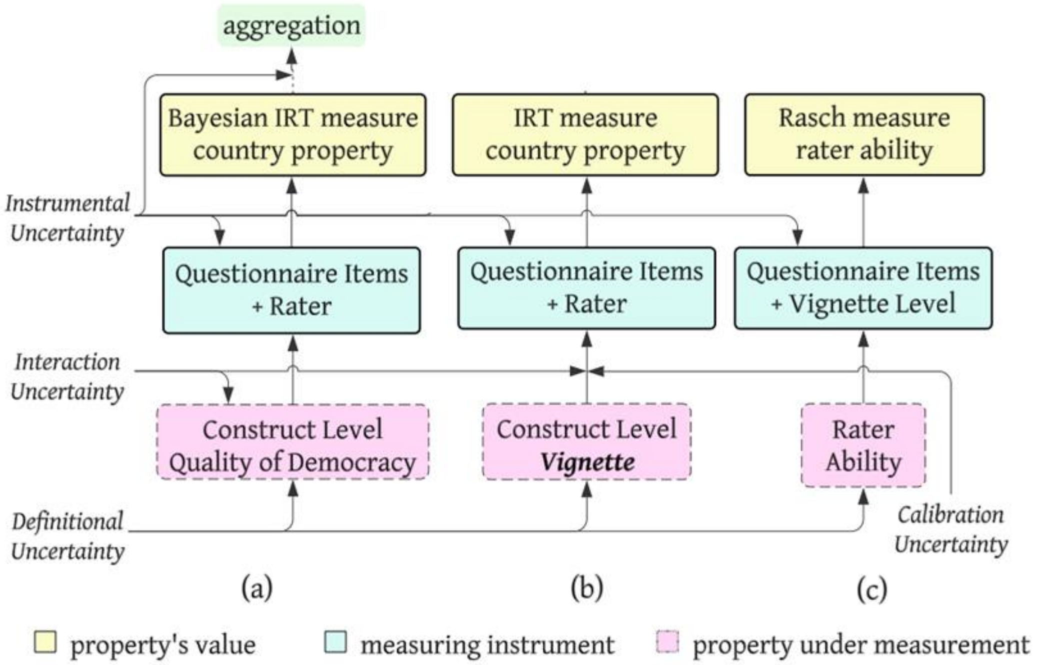
aggregation
Bayesian IRT measure country property
IRT measure country property
Rasch measure rater ability
Instrumental Uncertainty
Questionnaire Items + Rater
Questionnaire Items + Rater
Questionnaire Items + Vignette Level
Interaction Uncertainty
Construct Level Quality of Democracy
Construct Level Vignette
Rater Ability
Definitional Uncertainty
Calibration Uncertainty
(a)
(b)
(c)
property's value
measuring instrument
property under measurement

Figure 4.1. Diagrams of the measuring instrument models associated: in (a), with the conventional application of the V-Dem in measuring the quality of democracy in a country; in (b), with V-Dem's use of anchoring vignettes as quality-level references; and, in (c), with the proposed model focused on assessing the coders' ability to identify the degree of a country's democracy attribute using anchoring vignettes.

The V-Dem measurement model includes a scoring element, which corresponds to its sensing device, that involves at least five coders. These coders use a set of items to assess each property feature, ultimately providing a comprehensive assessment of a country's democracy quality (**Figure 4.1a**). Each item in the V-Dem measurement model is designed to probe the level of a specific attribute of a primary construct. This process is performed by coders, whose expertise and judgment are crucial for the accuracy of the measurement results.

The potential ambiguity or lack of clarity in the definition of the property to be measured can affect the accuracy of the measurement by contributing to definitional uncertainty (**Figure 4.1**), one of the main sources of uncertainty in political science research.

Moreover, the property's manifestation (country-specific information) or description (vignette) can influence the interpretation of the coders, who act as sensing elements, thus contributing as a source of interaction uncertainty (**Figure**

4.1a) or calibration uncertainty (**Figure 4.1b**). In addition to this interaction of the sensing elements with the information about the latent trait level in a given context, raters are also affected by the interaction with the text of the questionnaire items and their response options. These elements can be misinterpreted, affecting the correspondence between the level of the construct being measured and the scale defined in the measuring instrument [13] and contributing as a source of instrumental uncertainty (**Figure 4.1a** and **Figure 4.1b**).

The result obtained is thereby influenced by these different sources of uncertainty, aggravated by the fact that the sensor element consists of a human rater, which implies the use of a new sensor for each operation in a series of repeated measurements. The variety of metrological characteristics of these sensor elements can thus contribute as a source of instrumental uncertainty (**Figure 4.1a** and **Figure 4.1b**).

All these influences may generate additional quantities that affect the reading of the construct level and the interpretation of the item, potentially compromising the local independence and unidimensionality of the measurement process.

To develop a method for the metrological characterisation of the measurement sensor, this study used a set of V-Dem anchoring vignettes to act as a measurement standard, i.e. a reference material. In this case, the evaluation was represented by a measuring system for assessing the transducer performance, consisting of both the indicators (questionnaire items) and the vignettes, as shown in **Figure 4.1c**. The set of vignettes and items acted as a stimulus for the raters to provide responses that could be analysed using the standard values assigned to the vignettes (**Figure 4.1c**). This procedure can reveal possible biases (systematic errors).

For evaluating the sensing elements, measurement error was examined based on the raters' success or failure in identifying the construct level represented by each vignette. To this end, raters' responses to the two extreme construct levels of the vignettes associated with each of the seven items in the V-Dem Deliberation survey were transformed into dichotomous scores by converting the originally polytomous scores. Based on the resulting scores and using the Rasch measure, the coders' performance was modelled in terms of the difference between the respondents' ability to classify the vignettes as expected and the difficulty of each indicator-vignette combination in eliciting the intended responses.

Local contextual factors hindering access to reliable information about the assessed concept may affect the coders' ability to determine the property accurately. Similarly, in the context of vignettes, how coders interact with the synthetic reference text may affect the recognition of the underlying construct level. Therefore, as shown in **Figure 4.1**, the measurement results are affected by influence factors associated with definitional, interaction and instrumental uncertainty; these two latter components taking into account raters' attributes such as their ability, their severity and other elements that affect their interpretation given their different contexts of origin.

In the proposed measurement instrument model (**Figure 4.1c**), the influence of the rater's interaction with the questionnaire items and related responses' options is no longer an element of instrumental uncertainty but constitutes an interaction source of uncertainty, alongside the impact of the rater's interaction with the description of the vignette's construct level.

4.4 Results from the proposed approach

Figure 4.2 shows the Rasch estimates (in logit) for each of the fourteen items (indicator-vignette combinations), considering the difficulty they impose on raters to adequately categorise the vignette's construct level. These results derive from the proposed approach (**Figure 4.1c**) and consider the three main sample groups described in section 4.1: raters' ratings of the minimum-level vignettes (MIN group), ratings of the maximum-level vignettes (MAX group), and ratings of both groups of vignettes (MAX-MIN group).

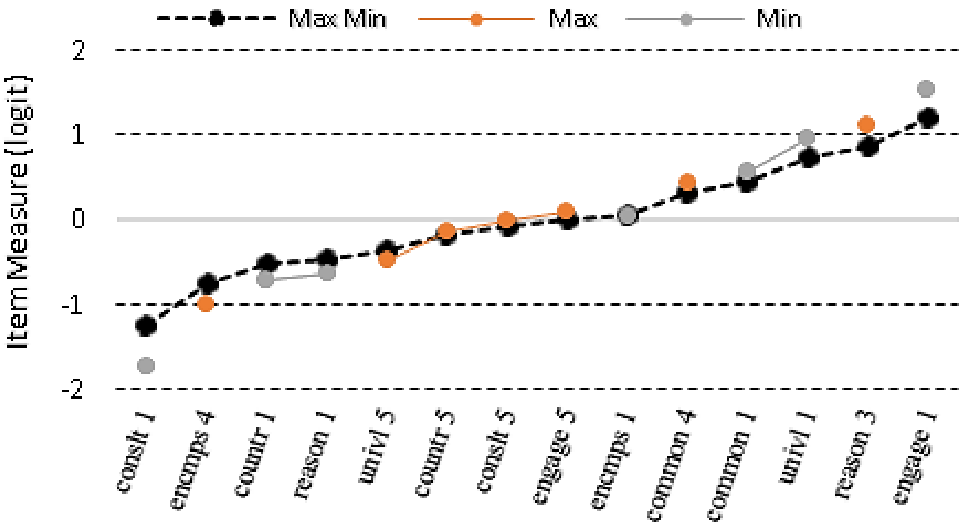

Figure 4.2. Difficulty measures for the fourteen indicator-vignette combinations across the three samples of respondents analysed (MAX-MIN, MAX, MIN). Indicator-vignette combinations are labelled according to Table 4.2, and arranged from left to right in increasing order of difficulty.

Based on the estimated item difficulties and considering, in particular, the results for the MAX-MIN sample, **Table 4.5** shows the probabilities of coders detecting the vignettes' predefined construct level on each of the fourteen indicator-vignette pairs, according to the measure of coders' ability (in logits). This indication follows the relationship formalized in equation (1), pertaining to the Rasch model. By displaying the probabilities of item responses as one progresses along the scale, **Table 4.5** facilitates the interpretation of the results, suggesting potential explanations for the underlying variable of the resulting measurement scale. According to **Table 4.5**, for instance, coders with ability measure estimated as -1.25 logit (i.e., 1.25 logit below the average item difficulty) are 50 % likely to rate 'correctly' the vignette anchored at the lowest level of the 'range of consultation' indicator ('conslt 1', with difficulty of -1.25 logit) but are less likely to 'correctly' rate any of the other vignettes. Meanwhile, at the other end of the scale, a coder with ability level of +1.21 logit is 50 % likely to 'correctly' rate the vignette representing the lowest level of the 'engaged society' indicator ('engage 1', with difficulty of 1.21 logit), and even more likely to do so for any of the other vignettes.

Table 4.5. Probability of 'correct' rating to each indicator-vignette pair according to coder ability.

item difficulty (logit)	-1.25	-0.75	-0.53	-0.48	-0.37	-0.18	-0.09	0.01	0.05	0.32	0.45	0.74	0.87	1.21
Probability of correct rating to each item according to coder ability														
Ability level (logit)	conslt 1	encmps 4	countr 1	reason 1	univl 5	countr 5	conslt 5	engage 5	encmps 1	common 4	common 1	univl 1	reason 3	engage 1
1.21	92 %	88 %	85 %	84 %	83 %	80 %	79 %	77 %	76 %	71 %	68 %	62 %	58 %	**50 %**
0.87	89 %	83 %	80 %	79 %	78 %	74 %	72 %	70 %	69 %	63 %	60 %	53 %	**50 %**	42 %
0.74	88 %	82 %	78 %	77 %	75 %	72 %	70 %	67 %	67 %	60 %	57 %	**50 %**	47 %	38 %
0.45	85 %	77 %	73 %	72 %	69 %	65 %	63 %	61 %	60 %	53 %	**50 %**	43 %	40 %	32 %
0.32	83 %	74 %	70 %	69 %	67 %	62 %	60 %	58 %	57 %	**50 %**	47 %	40 %	37 %	29 %
0.05	79 %	69 %	64 %	63 %	60 %	56 %	53 %	51 %	**50 %**	43 %	40 %	33 %	31 %	24 %
0.01	78 %	68 %	63 %	62 %	59 %	55 %	52 %	**50 %**	49 %	42 %	39 %	33 %	30 %	23 %
-0.09	76 %	66 %	61 %	60 %	57 %	52 %	**50 %**	48 %	47 %	40 %	37 %	30 %	28 %	21 %
-0.18	74 %	64 %	59 %	57 %	55 %	**50 %**	48 %	45 %	44 %	38 %	35 %	28 %	26 %	20 %
-0.37	71 %	59 %	54 %	53 %	**50 %**	45 %	43 %	41 %	40 %	33 %	31 %	25 %	22 %	17 %
-0.48	68 %	57 %	51 %	**50 %**	47 %	43 %	40 %	38 %	37 %	31 %	28 %	23 %	21 %	16 %
-0.53	67 %	55 %	**50 %**	49 %	46 %	41 %	39 %	37 %	36 %	30 %	27 %	22 %	20 %	15 %
-0.75	62 %	**50 %**	45 %	43 %	41 %	36 %	34 %	32 %	31 %	26 %	23 %	18 %	17 %	12 %
-1.25	**50 %**	38 %	33 %	32 %	29 %	26 %	24 %	22 %	21 %	17 %	15 %	12 %	11 %	8 %

It is worth noting that, when using the measuring instrument proposed in this study (**Figure 4.1c**), the difficulty level of an item is directly related to the performance of the coders, who are the sensing elements of the V-Dem measuring system. If applied to the conventional configuration of this system instead (**Figure 4.1a** and **Figure 4.1b**), Rasch measures of item difficulty would correspond to the level of the construct assessed by the indicators, meaning that a greater amount of that construct would be needed to meet the criteria set out by the more challenging indicators.

The difficulty measures reflecting raters' performance on each indicator-vignette combination was relatively the same for the three main samples of respondents analysed (**Figure 4.2**). The most difficult item comprised the minimum level of the construct assessed by the "engaged society" indicator (engage 1). Above the average item difficulty –fixed at 0 logit for each sample analysed–, this item was followed by 'reason 3', 'univl 1', 'common 1', 'common 4' and 'encmps 1' (**Figure 4.2** and **Table 4.5**). This indicates that, for four of the seven items (slightly more than half), it was more difficult for coders to identify the constructs at their lowest level. Likewise, considering the estimates from the MAX-MIN sample, the average difficulty of the minimum-level vignettes (0.027 logit) was slightly above the one for maximum-level ones (-0.027 logit), both equidistant from the average difficulty of the whole set of items. However, with the standard deviation of difficulty levels being around 0.8 logit across minimum-level vignettes and 0.5 across maximum-level ones, the difference, in less than 0.1 logit, between the average difficulties of these two subgroups is not statistically significant, with p-value=0.896 (>0.05).

By solely assessing vignettes representing the minimum and maximum latent trait levels, comparisons between each pair of vignettes for every indicator can unveil whether coders exhibited unexpected leniency or rigidity. The findings indicate a tendency for coders to lean towards greater rigour in evaluating the construct of ‘reasoned justification’, as its highest-level vignette (‘reason 3’) frequently received ratings falling into lower-level categories, whereas the reverse was not observed. Conversely, coders displayed notable leniency in scoring the 'engaged society' indicator, often assigning high scores to the lowest-level vignette (‘engage 1’) while not erroneously lowering their level indication for the maximum construct level vignette (‘engage 5’).

The appropriateness of raters' responses to the vignettes was designed in alignment with the V-Dem method for defining thresholds to classify a latent trait level [152]. It should be noted, however, that items did not share the same number of ordinal levels as response options, varying from 4 to 6 categories. The second most challenging item, corresponding to the maximum-level vignette for the construct of 'reasoned justification' (‘reason 3’), is the only item with just four response categories. In this case, an inadequate response means classifying the description of the highest level of the vignette's construct as the lowest. The higher item difficulty suggests that a particular aspect of the construct expression at its maximum level, or an item description when interpreted in conjunction with a high level of the construct, impacts a larger group of raters, hindering their appropriate interpretation and, consequently, the adequate discrimination of the latent trait level.

The difficulty measures of the fourteen items (indicator-vignette combinations) of the MAX-MIN sample are shown in **Table 4.6**, alongside their random error –represented by the the “Model Standard Error” ('*Model S.E.*')– and fit statistics –namely, the meansquare ('*Mnsq*') and standardised ('*Zstd*') values of the *outfit* and *infit* indices. The *total score*, representing the sum of correct responses observed for each item, is also indicated.

Table 4.6. Item statistics for the MAX-MIN sample, indicating the sum of correct responses (total score) for each of the fourteen items of the MAX-MIN sample, their Rasch measure of difficulty, random error (Model S.E.) and fit statistics –mean-square and standardised values of infit and outfit. The items (indicator-vignette combinations) are labelled according to Table 4.2.

Total Score	**Item Measure (logit)**	**Model S.E.** (logit)	**Infit**		**Outfit**		**items**	
			Mnsq	Zstd	Mnsq	Zstd	(indicator + vignette)	
72	1.21	0.21	1.4	4	1.52	3.6	engage	1
80	0.87	0.21	1.33	3.5	1.42	3.2	reason	3
83	0.74	0.21	1.07	0.9	1.04	0.4	univl	1
90	0.45	0.21	0.95	-0.5	0.96	-0.2	common	1
93	0.32	0.21	1.34	3.5	1.34	2.3	common	4
99	0.05	0.21	0.93	-0.7	0.92	-0.4	encmps	1
100	0.01	0.21	0.87	-1.4	0.86	-0.9	engage	5
102	-0.09	0.22	0.82	-2	0.77	-1.4	conslt	5
104	-0.18	0.22	0.68	-3.8	0.54	-3.1	countr	5
108	-0.37	0.22	0.91	-0.9	0.7	-1.6	univl	5
110	-0.48	0.23	1.11	1	1.04	0.3	reason	1
111	-0.53	0.23	0.84	-1.5	0.65	-1.8	countr	1
115	-0.75	0.24	0.89	-0.9	0.9	-0.3	encmps	4
123	-1.25	0.27	0.81	-1.2	0.64	-1.1	conslt	1
99.3	0	0.22	1	0	0.95	-0.1	**MEAN**	
13.8	0.65	0.02	0.22	2.2	0.29	1.9	**P. SD.**	

Table 4.6 indicates three items with data underfitting the Rasch model: 'engage 1', 'reason 3' and 'common 4'. These same items also exhibited the most underfitting response patterns when analysing the MAX and MIN samples separately. 'Engage 1', the most underfitting item, was the indicator-vignette stimulus with the highest difficulty measure (**Table 4.6**). Likewise, the other two underfitting items ('reason 3' and 'common 4') were the most difficult of the indicator-vignette stimuli associated with the highest construct level. The observed misfit suggests a random or at least non-uniform underlying basis for these three items with the highest difficulty level.

Rasch measures, random errors, and goodness of fit statistics, for persons (raters) and items (indicator-vignette combinations), are summarised in a single graph, as proposed by [4], and presented in **Figure 4.3**, **Figure 4.4**, and **Figure 4.5,** for each of the three samples of responses analysed: MAX-MIN, MIN, and MAX, respectively. The vertical axis indicates the Rasch measures are presented in a logit scale, along which person abilities and item difficulties are located, while the horizontal axis informs the standardised values of infit. Persons and items are plotted as circles, with the diameter representing the size of random error associated with each of their measures in the logit scale (in the vertical axis). Each individual person is plotted as a semi-transparent black circle, so darker circles indicate a higher concentration of persons with that particular measure and fit values. The remaining circles represent the items (indicator-vignette combinations), with pink circles associated with minimum-level vignettes and blue circles associated with maximum-level ones.

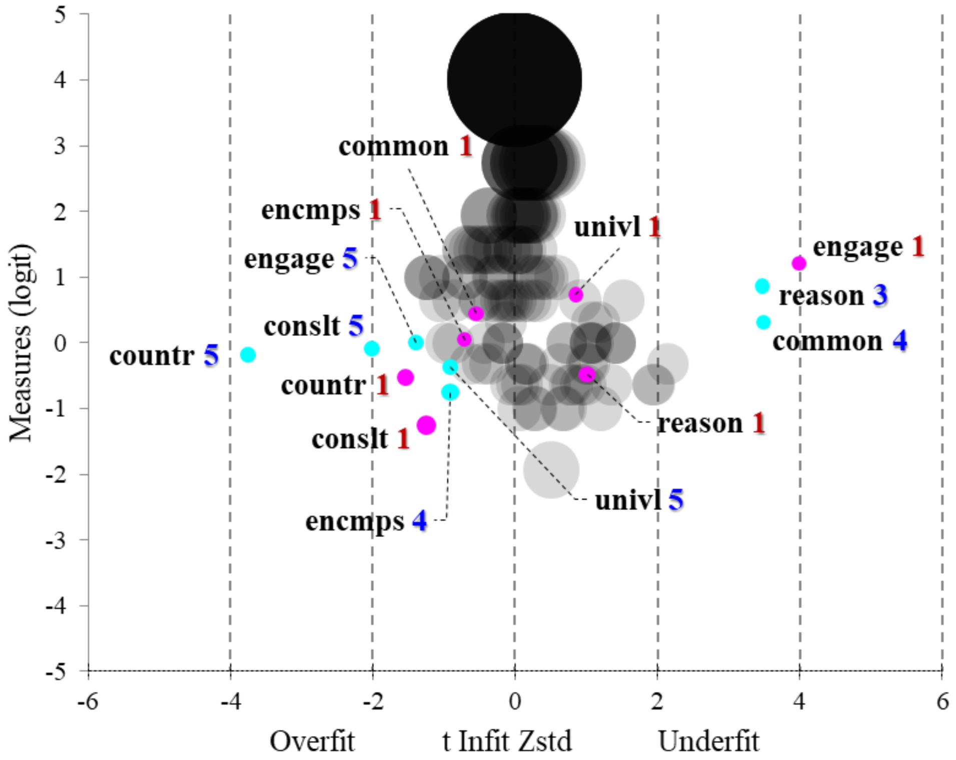

Figure 4.3. Raters' and items' positions according to their Rasch measure and standardised infit value [4] for the MAX-MIN sample. The semi-transparent black circles represent individual coders, with darker regions indicating a more significant number of coders plotted in the same area. Pink circles

display minimum-level vignettes, while blue circles indicate maximum-level vignettes. The diameter of each circle indicated along the vertical axis (in logit) denotes the random error associated with the corresponding Rasch measure.

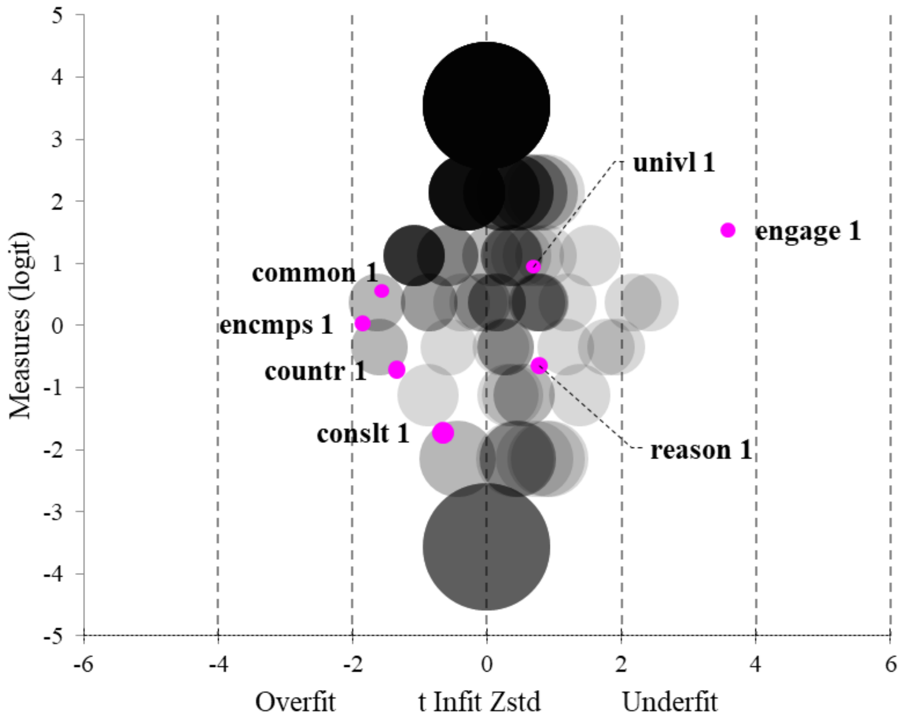

Figure 4.4. Raters' and items' positions according to their Rasch measures and standardised infit values [4] for the MIN sample. Pink circles indicate minimum-level indicator-vignettes. See Figure 4.3 caption for a complete description of the illustrated elements.

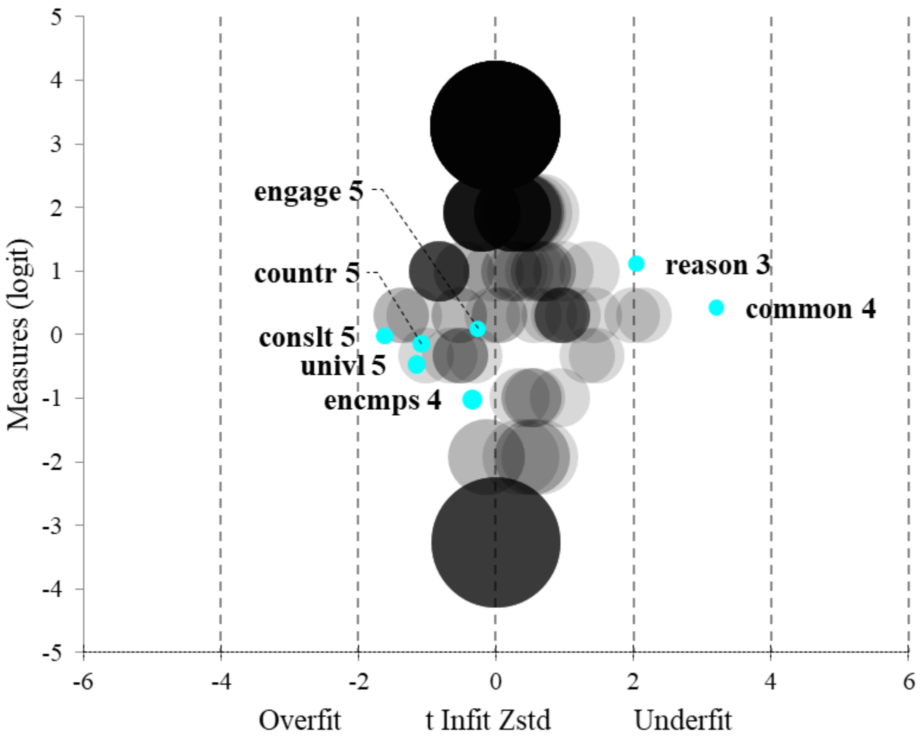

Figure 4.5. Raters' and items' positions according to their Rasch measures and standardised infit values [4] for the MAX sample. Blue circles indicate maximum-level indicator-vignettes. See Figure 4.3 caption for a complete description of the illustrated elements.

The reliability and separation indices based on the Rasch model, which refer to the proportion of variance that can be associated with the construct being measured [4], [10], [68], [105], [106], are presented in **Table 4.7**, for both item and person parameters.

Table 4.7. Rasch model reliability and separation indices for person and item measures, calculated from each sample of responses (MIN, MAX, MAX MIN)

sample		item	person
MIN	reliability	0.94	0.49
	separation index	3.83	0.99
MAX	reliability	0.83	0.45
	separation index	2.18	0.91
MAX-MIN	reliability	0.87	0.61
	separation index	2.64	1.24

With item difficulty measures covering only about half the range of raters' abilities (**Figure 4.3** and **Figure 4.4**) –or a third of it, in the case of the MAX sample (**Figure 4.5**)–, higher uncertainty (**Figure 4.3**, **Figure 4.4**, and **Figure 4.5**) and lower reliability estimates (**Table 4.7**) were observed for person measures (rater ability) compared to those of the items. The random error for item measures remained around 0.25 logit in all three samples, while, for person measures, it ranged from more than 0.5 to almost 2 logits. Nonetheless, given the approach adopted in this analysis (**Figure 4.1c**), the absence of items that match the full spectrum of raters' abilities is not inherently problematic as it does not indicate a flaw in the rating system.

As evidenced in **Figure 4.3**, most coders' ability levels exceed the item difficulties set, suggesting that coders were generally successful in identifying the construct level represented in each anchoring vignette. However, it should be noted that, in the conventional application of the V-Dem measuring system (**Figure 4.1a**), each country is evaluated by a group of five experts rather than the entire set of raters considered in this study. Therefore, the appropriateness of the instrument, considering the adequacy of raters' interpretation, must be ensured for every small group of coders.

Figure 4.6, **Figure 4.7** and **Figure 4.8**, pertaining to each of the three major data samples (MAX-MIN, MIN, and MAX, respectively), show the item difficulty measures estimated from the three largest subsamples of coders grouped by continent –A, C, and D, each with over 30 coders, as detailed in **Table 4.3**. The items' difficulty measures based on the responses from the main data sample (MAX-MIN, MIN, or MAX) are also included in each figure.

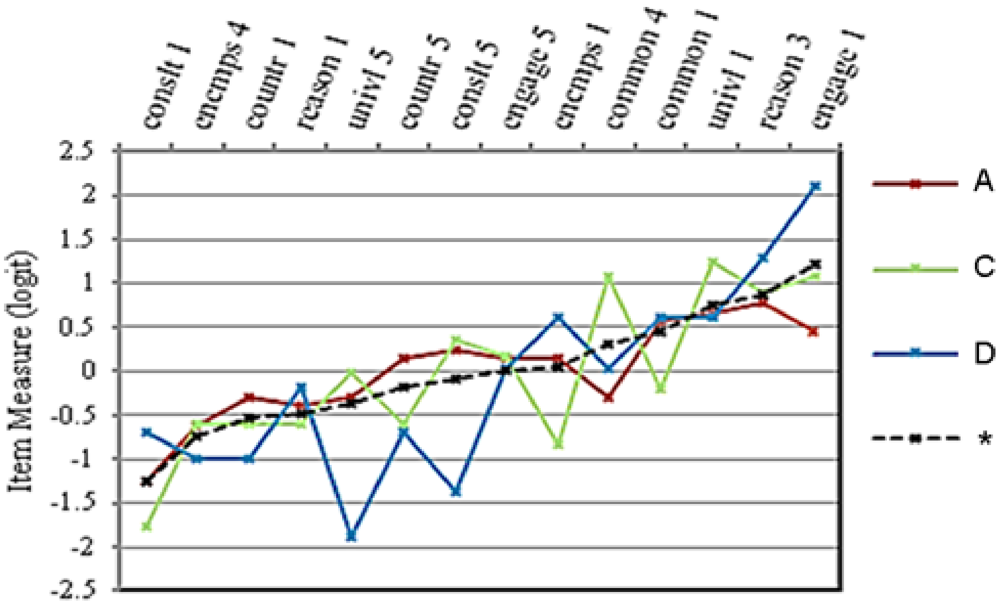

Figure 4.6. Item difficulty measures estimated based on the three largest subsamples of coders by continent (A, C and D) in the MAX-MIN sample, compared with the item measures based on the full MAX-MIN sample (*), shown in Figure 4.2.

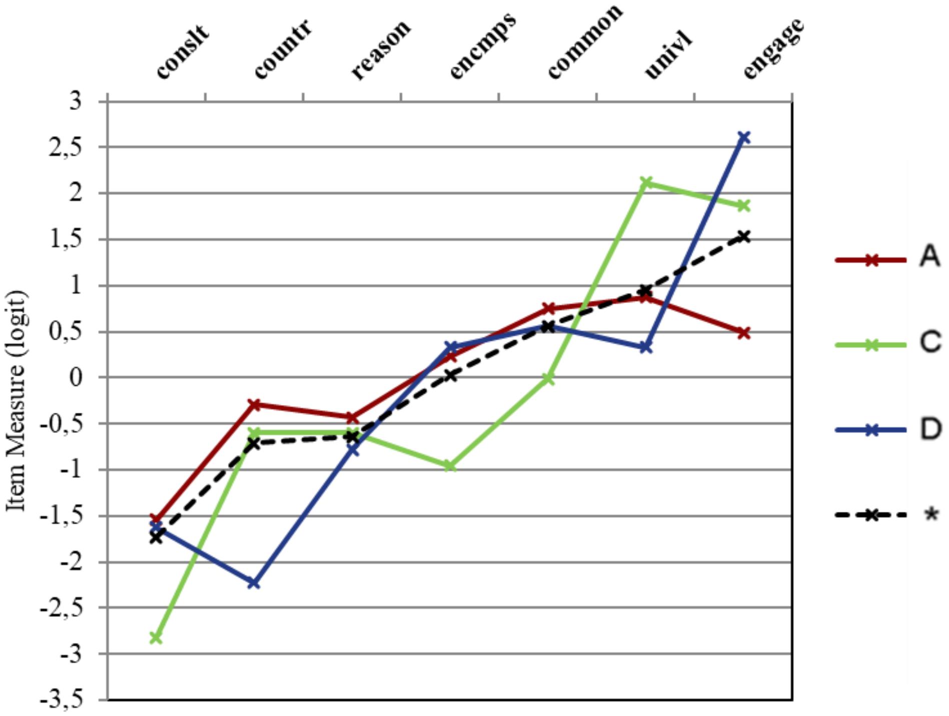

Figure 4.7. Item difficulty measures estimated based on the three largest subsamples of coders by continent (A, C and D) in the MIN sample, compared with the item measures based on the full MIN sample (*), shown in Figure 4.2.

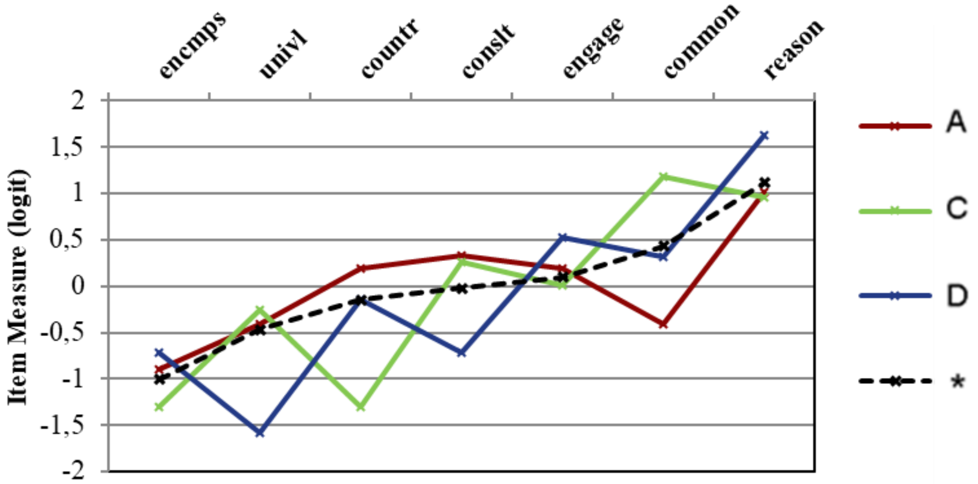

Figure 4.8. Item difficulty measures estimated based on the three largest subsamples of coders by continent (A, C and D) in the MAX sample, compared with the item measures based on the full MAX sample (*), shown in Figure 4.2.

Differential item functioning (DIF) was observed to have statistical significance (p≤0.05) for some of the most difficult items. 'Engage 1', which was the most difficult indicator-vignette stimulus (**Figure 4.2**), and had the most underfitting pattern of coders' performances (**Figure 4.3**), is the only item that showed significant DIF in both samples in which it was included (MIN and MAX-MIN). Coders of group D performed worse than expected on this item, with a DIF size of +0.9 logit in the MAX-MIN sample (**Figure 4.6**) and of +1.07 logit in the MIN (**Figure 4.7**). Group A, in contrast, performed better than expected on this item, with a DIF size of -0.76 logit in the MAX-MIN sample (**Figure 4.6**) and -1.05 logit in the MIN sample (**Figure 4.7**). For this group, unlike the other subsamples of coders, 'engage 1' was not the most difficult item, ranking behind 'reason 3', 'univl 1', and 'common 1' (**Figure 4.6**).

When responses to maximum level vignettes are considered in isolation (MAX), coders from group A performed better than expected at identifying the highest level of 'Common good' (common 4), with the difficulty of this item being 0.83 logit lower than when estimated considering the whole group of MAX sample coders –i.e., presenting a DIF size of -0.83 logit (**Figure 4.8**). When considering separately the responses to minimum-level vignettes (MIN), in turn, in addition to the unexpected performance in 'engage 1', 'univl 1' showed significant signs of DIF, in which coders from group C performed worse than predicted by the model, with a DIF size of +1.17 (**Figure 4.7**).

To evaluate if the responses to the instrument's stimuli have a one-dimensional structure, a Principal Component Analysis of Rasch model residuals (PCAR) was conducted [4], [10], [105], [106], [153], [156], [157], [159], [160]. Since it is performed on the residuals and not on the original data, any factor identified by this analysis –constituting the PCA contrasts– pertains to a (possible) secondary dimension other than the one measured by the model [153]. The raw variance explained by measures and the unexplained variance found in the first PCA contrast of the residuals from the data pertaining to each of the three main samples analysed (MAX-MIN, MAX, and MIN) are indicated in **Table 4.8**.

Table 4.8. The variance explained by measures and the unexplained variance in the first PCA contrast of residuals, for each of the three main data samples used in this study (MAX-MIN, MAX, and MIN)

	MAX-MIN	**MAX**	**MIN**
Variance explained by measures (eigenvalues)	4.5	2.3	3.7
Residual variance in the 1st contrast (eigenvalues)	2.8	2	1.5

Considering the performances of the expert coders in evaluating the minimum and maximum level vignettes for each of the seven selected V-Dem indicators (MAX-MIN sample), the residuals' variance explained by the first PCA contrast corresponded to approximately 2.8 eigenvalue units (**Table 4.8**), suggesting that the performances in at least two items could be related to a secondary dimension other than that measured by the model. Nevertheless, the variance explained by measures was around 4.5 eigenvalues, almost twice the amount attributed to the residuals in the first contrast (**Table 4.8**).

When the responses to the maximum and minimum-level vignettes are considered separately, the residual variance in the first contrast decreases to around 2 eigenvalues for the MAX sample and 1.5 eigenvalues for the MIN sample (**Table 4.8**). In the MAX sample, the amount of residual variance in the first contrast is about the same as that explained by measures (in around 2.3 eigenvalues), while, in the MIN sample, the model explained variance (3.7 eigenvalues) is more than twice the one attributed to the first contrast of item residuals (**Table 4.8**). Therefore, the possibility of a secondary dimension could be mainly associated with the indicator-vignette maximum level pairs.

Figure 4.9, **Figure 4.10** and **Figure 4.11** show each item's residual loadings in the first contrast of the PCAR for the data comprising the MAX, MIN, and MAX-MIN samples, respectively. The items are distributed horizontally according to their difficulty and vertically according to their residual loadings in the first contrast. For each sample, the items are also grouped in three clusters according to their residual loadings in the first PCAR contrast, as shown in **Table 4.9**.

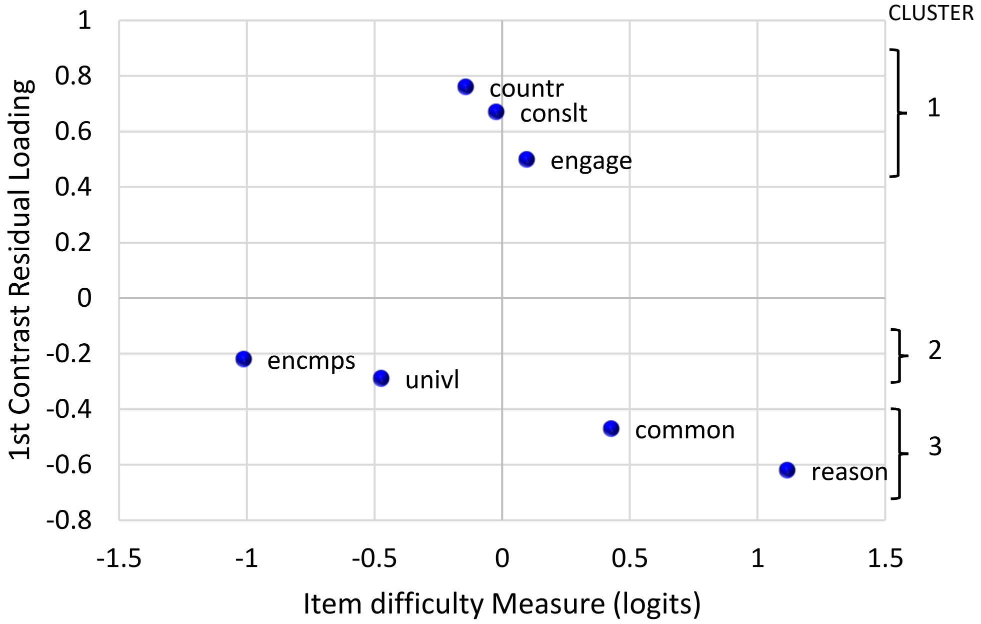

Figure 4.9. Items from the MAX sample ranked according to their difficulty measures (horizontal axis) and their residual loadings in the first PCAR contrast (vertical axis). On the right side of the graph, the items are grouped in three clusters according to their residual loadings in the first contrast.

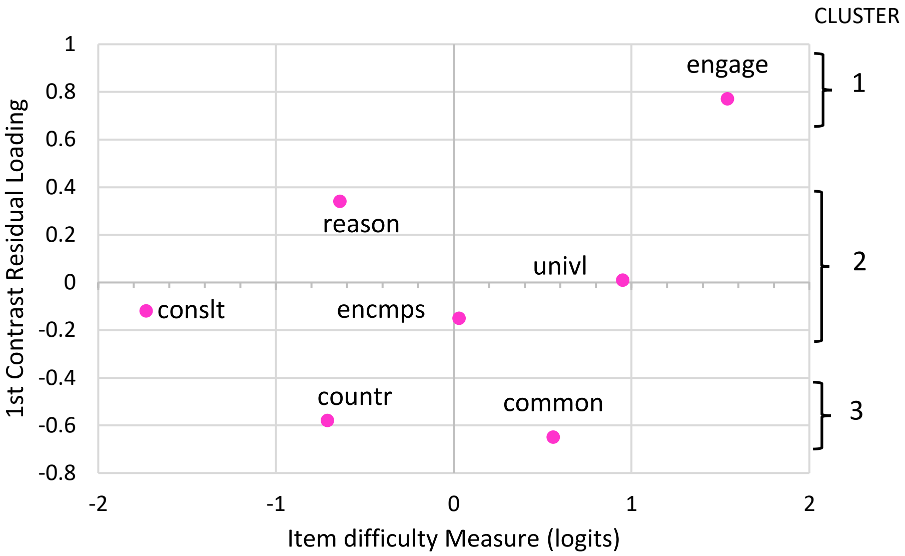

Figure 4.10. Items from the MIN sample ranked according to their difficulty measures (horizontal axis) and their residual loadings in the first PCAR contrast (vertical axis). On the right side of the graph, the items are grouped in three clusters according to their residual loadings in the first contrast.

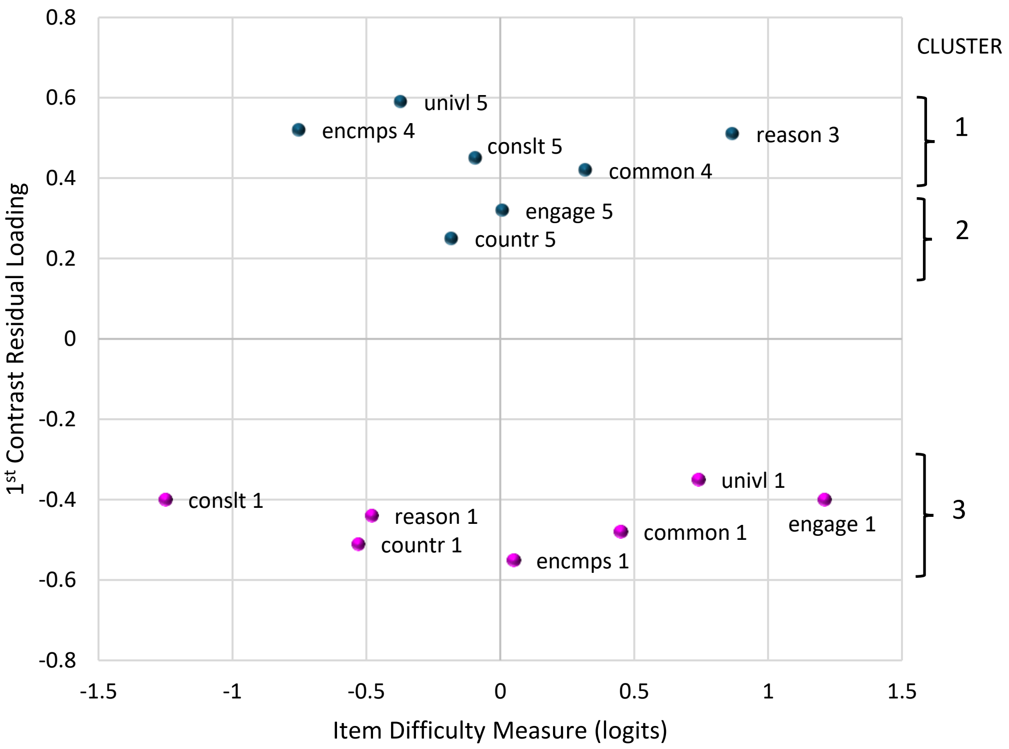

Figure 4.11. Items from the MAX-MIN sample ranked according to their difficulty measures (horizontal axis) and their residual loadings in the first PCAR contrast (vertical axis). On the right side of the graph, the items are

grouped in three clusters according to their residual loadings in the first contrast.

Table 4.9. Items grouped by clusters according to their residual loadings in the first PCAR contrast, for each of the three main samples (MAX, MIN and MAX-MIN) –as shown in Figure 4.9, Figure 4.10 and Figure 4.11.

MAX			**MIN**		
item	loading	cluster	item	loading	cluster
countr 5	0.76	1	*engage 1*	0.77	1
conslt 5	0.67	1	*reason 1*	0.34	2
engage 5	0.5	1	*univl 1*	0.01	2
encmps 4	-0.22	2	*conslt 1*	-0.12	2
univl 5	-0.29	2	*encmps 1*	-0.15	2
common 4	-0.47	3	*countr 1*	-0.58	3
reason 3	-0.62	3	*common 1*	-0.65	3

MAX-MIN					
item	loading	cluster	item	loading	cluster
univl 5	0.59	1	*univl 1*	-0.35	3
encmps 4	0.52	1	*conslt 1*	-0.4	3
reason 3	0.51	1	*engage 1*	-0.4	3
conslt 5	0.45	1	*reason 1*	-0.44	3
common 4	0.42	1	*common 1*	-0.48	3
engage 5	0.32	2	*countr 1*	-0.51	3
countr 5	0.25	2	*encmps 1*	-0.55	3

Figure 4.11 displays a clear separation between maximum and minimum-level vignettes, with the former exhibiting only positive residual loadings in the first contrast and the latter only negative ones. More specifically, five items comprising maximum-level vignettes were identified at the top of the plot (cluster 1), with contrast loadings larger than 0.4, while all seven items associated with minimum-level vignettes were identified at the bottom (cluster 3), with contrast loadings close to -0.4 or lower (**Table 4.9**). These two clusters entail items whose residual patterns deviate to some extent from the patterns expected by the measurement model.

To further examine the invariance of estimates and thereby the dimensionality of the responses to the measuring instrument, the correlation between each rater's ability estimated by the different item clusters identified in the PCAR (**Table 4.9**) is shown in **Table 4.10**, for each of the three major data samples (MAX, MIN, and

MAX-MIN). Along with the Pearson correlation, the table includes the disattenuated value of the correlation, which takes into account the random error associated with the person measures produced by each item cluster [161]. Measures of raters with extreme performances –i.e., who succeeded or failed in identifying the construct levels of all indicator-vignettes within the selected sample (MIN, MAX, or MAX-MIN)–, featuring great uncertainty [4], are not included in any of these correlation indices.

Table 4.10. correlation of rater ability measures across the three clusters of items grouped by their residual loadings in the 1st PCAR contrast, for each of the three main data samples (MIN, MAX, and MAX-MIN)

sample	Clusters	Pearson correlation	Disatenuated correlation
MIN	**1 - 2**	-0.12	-1
	2 - 3	0.45	1
	1 - 3	-0.15	-1
MAX	**1 - 2**	0.22	1
	2 - 3	0.1	1
	1 - 3	-0.27	-1
MAX-MIN	**1 - 2**	0.49	1
	2 - 3	0.27	1
	1 - 3	-0.1	-0.24

As shown in **Table 4.10**, the measures of rater ability produced by the two most opposing item clusters (1 and 3) from the MAX-MIN sample –contrasting items associated with minimum level vignettes with those related to maximum level vignettes (**Figure 4.11**)– were uncorrelated.

On the other hand, the correlation between the ability measures derived from cluster 2 ('engage 5' and 'countr 5') and those derived from either of the other two item clusters (1 or 3) is higher than the correlation between clusters 1 and 3 (**Table 4.10**). The Pearson correlation of the rater ability measures was of approximately 0.49 for clusters 1 and 2, and 0.27 for clusters 2 and 3. The disattenuated correlation value is considerably elevated (**Table 4.10**) due to the high random error attributed to the person measures in item cluster 2, formed by two items only—which are too few to assess the performance of the 142 raters adequately.

Considering the item residuals in the first PCAR contrast for the MIN data sample, cluster 1 was represented by a single item, 'engage 1', while two items constituted cluster 3: 'common 1' and 'countr 1' (**Figure 4.10** and **Table 4.9**). Hence, most items were grouped in cluster 2, with no significant residual loadings in the first contrast. This outcome may be associated with the broader range of variance explained by measures (3.7 eigenvalues), in the MIN sample, compared to the variance explained by the first contrast of the residuals (1.5 eigenvalue), as shown in **Table 4.8**.

The item for detecting the lowest level of the ‘engaged society’ indicator (‘engage 1’), which was the single component of cluster 1 in the MIN sample, was also the item with the highest difficulty and underfit to the Rasch model (**Figure 4.3**, **Figure 4.4**, and **Table 4.6**). The Pearson correlation between raters' ability based on item cluster 1 ('engage 1') and either of the two other item clusters was particularly low (around -0.12 and -0.15), compared to the correlation between clusters 2 and 3 (0.45). The disattenuated correlation, which compensates for the random error in person measures, was nevertheless considerable for all three combinations of item clusters (**Table 4.10**). Still, reflecting the values for the Pearson correlation, positive correlation was observed only between clusters 2 and 3 (**Table 4.10**).

When examining the MAX sample, 'reason 3' and 'common 4', both the least fitting and most difficult items of this sample, were observed to have the most negative loadings in the first contrast, forming item cluster 3 (**Figure 4.9** and **Table 4.9**). On the opposite side, in cluster 1, the three other indicators of the V-Dem Deliberation survey that were also comprised in the V-Dem index of the deliberative component of democracy [24] –'countr 5', 'conslt 5' and 'engage 5'– were the only items with positive residual loadings in the first contrast (**Figure 4.9** and **Table 4.9**). The correlation of person measures between either of these two most opposing item clusters and cluster 2 was low (**Table 4.10**), which could suggest the possibility of one or more secondary dimensions affecting the sensors on their responses to those items. Nevertheless, with high random errors associated with the measures estimated based on each of these item clusters (with clusters 2 and 3 comprising only two items each), the disattenuated correlation for all three cluster combinations was considerable (**Table 4.10**).

Correlation analysis of the item residuals to identify local dependence between item responses shows a loss of independence only when the joint MAX-MIN sample is used. Given that each coder makes two entries in the dataset for the MAX-MIN sample, the performance of the same coder on the same item in different situations may have led to local dependence. Local dependence between 'conslt 5' and 'countr 5' (standardised residual correlation of 0.39) is probably related to the overfit behaviour of these items, as shown in **Table 4.6** (standardised infit ≤-2). A slight correlation of residuals in the first contrast is also observed between the two items of the dataset that are associated with V-Dem’s egalitarian democracy index ('encmps 4', 'univl 5'), possibly indicating that a common element may be needed for the raters to identify these constructs at their highest levels. As shown in **Table 4.2**, these two items, both involving raters' assessment of aspects of public policy related to the egalitarian principle of democracy [24], showed similar residual loadings in the first PCAR contrast.

5
Discussion, Conclusions and Future Work

The literature explored in Chapter 2, the results of which are published in [14], pointed to concerns regarding the application of metrological principles to the study of social phenomena that have emerged since sociology's early development as a distinct discipline. Max Weber's concept of Ideal Type, which closely resembles the use of 'reference materials' in chemistry and biology, is an early example. As noted, Weber's Ideal Types provided a foundation for later advancements in psychometrics –notably, Louis Guttman’s scale approach, which assesses the data fit to the principles of measurement invariance. Georg Rasch further developed this concept probabilistically.

The Rasch analysis allows psychosocial measurement to meet the same requirements as physical measurement. Recent studies in Measurement Science have recognized the Rasch Measurement Theory as ideal for enabling metrological traceability in the Social Sciences [8], [9], [14], [15], [46]-[49], [59]. This would involve creating item banks and property reference scales combined with Rasch model fitting [8], [13], [14].

In the field of democracy measurement, there is limited explicit mention of metrology [15]. However, the ongoing discussions surrounding current methods used to measure democracy, as detailed in Chapter 3 and published in [15], incorporate key elements of the Science of Measurement. Numerous studies on democracy measurement have highlighted issues with comparability, and scholars have expressed concerns regarding random error and systematic bias.

Various democracy measuring systems are available nowadays, and multiple approaches have been adopted to evaluate and enhance the quality of their results. In addition to case studies, which typically employ qualitative approaches, strategies for validating democracy measurements range from simple statistical techniques to more sophisticated models, including psychometric tools like Item Response Theory (IRT) and the Guttman scale analysis. However, none of these approaches enables the provision of a common scale with a fixed reference unit to ensure the comparability of democracy measurement results.

Hence, despite progress, there are still significant gaps in aligning the methods used in democracy measurement literature with fundamental metrological principles. Strategies linked to the Rasch measurement theory, which has shown promise in addressing metrological challenges in the human and social sciences, have been largely absent in democracy measurement [15], [85], [149]. A notable exception is [149], a recent study that applied the Rasch model to test the extent to which the underlying political trust scale is linear and hierarchical using data from the Democracy Barometer measuring system. Based on seven cross-national data sets from 161 national surveys applied in 119 countries and territories, the analysis performed in [149] aimed to provide a global perspective on the political trust items. Despite the partial scalar invariance being strongly considered in the measurement equivalence literature and the strong evidence for monotonous homogeneity in studies applying IRT scale analysis to these political trust items, [149] reported no evidence that the political trust items meet the demands for unidimensionality. Of the 161 surveys, Rasch model fitting was observed in only one, suggesting that there may be different dimensions associated with the various objects of trust (such as government, parliament, etc.), with unique meanings to the respondents, that extend beyond the notion of indicators on a single scale [149]: i.e., these differences were not mere differences in trust levels, but considered a variance structure relative to different political institutions.

The final chapters of this thesis describe the development and results obtained from a pilot study carried out by applying the Rasch model for the metrological characterisation of a democracy measuring system: the “Varieties of Democracy” (V-Dem) project. This study focused on analysing the “sensor elements” of the measuring system, constituted by human evaluators (coders), to identify potential issues that could affect the comparability of the measurement results.

The analysis ranged over the coders' responses to a complete V-Dem survey (seven indicators), the "Deliberation" survey, using "anchoring vignettes"—which are synthetic reference texts designed to represent specific thresholds on each indicator’s response scale. The study concentrated on the sharper contrast associated with rating vignettes describing hypothetical cases expected to lie at either the upper or lower extreme levels of each indicator's response scale. By treating the evaluators as the "sensors" and the vignettes as reference materials, this approach aimed to assess the metrological properties of the V-Dem measuring system. Rather than focusing on the democracy characteristics evaluated by the indicators, the analysis centred on the evaluators' abilities to categorize the vignettes according to their predefined construct levels. The said evaluators' ability was then the measurand for the proposed measuring instrument model (**Figure 4.1c**).

The coders' originally polytomous responses were converted into dichotomous scores, indicating success or failure in identifying the construct level of a given vignette. Based on these scores and using the Rasch probabilistic model of measurement, the coders' performance was modelled in terms of the difference between the respondents' ability to classify the vignettes as anticipated by the measuring instrument’s design and the difficulty of each indicator-vignette combination in eliciting the intended responses.

For both the proposed and conventional modelling of the V-Dem measuring system, possible sources of measurement uncertainty and their location in the measurement process were discussed. These sources of uncertainty include: (1) the ambiguity in defining the properties to be measured, which leads to definitional uncertainty; (2) the interaction between coders and the questionnaire items, which contributes to instrumental uncertainty in the conventional model, and interaction uncertainty in the proposed model, (3) the interaction between coders and the country-specific information or the vignette description, contributing to interaction uncertainty or, in the case of vignettes, in the conventional configuration of the measuring system, as a source of calibration uncertainty; (4) the variability in coders’ characteristics, and their interaction with other elements of the measuring system, constituting potential sources of instrumental uncertainty in the conventional model of the measuring system. The employed computation procedure may also contribute as a source of instrumental uncertainty.

By targeting raters’ ability as the measurand, the assessment based on the proposed approach explored the hypothesis of a single dimension underlying raters' success or failure in detecting the construct levels hypothetically associated with each measurement standard (anchoring vignette).

Evaluations revealed that, for the lowest construct level, the single indicator-vignette pair with responses significantly different from the expectations of the

Rasch measurement model was 'engage 1', which was also the most difficult for coders to identify correctly (**Figure 4.2** and **Figure 4.3**).

Dimensionality issues and underfit observed for 'engage 1' may have been caused by its Differential Item Functioning (DIF), with coders from group A (as distinguished by continent in **Table 4.3**) performing better than expected on this indicator-vignette pair and group D performing worse (**Figure 4.6**). Thus, group A coders were likely more rigorous than expected when evaluating this item, while group D coders were likely more lenient. The hypothesis relating the underfit in 'engage 1' with the DIF observed is further suggested by **Figure 5.1** and **Figure 5.2**.

Figure 5.1 presents the distribution of rater ability for each group of coders by continent, taken from the MAX-MIN sample. The approximate position of 'engage 1' along the scale, based on this item’s estimate of difficulty in the MAX-MIN sample, is also indicated in the figure.

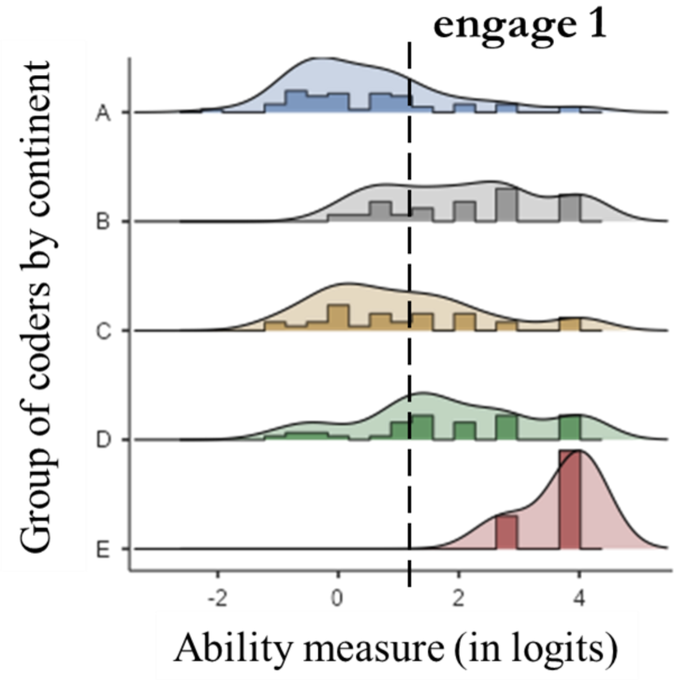

Figure 5.1. Distribution of raters as to their ability measures, in logit (horizontal axis), for each group of raters distinguished by their continents of origin (vertical axis), in the MAX-MIN sample; and the approximate position of item 'engage 1' (indicated by the dashed line) in that same scale (horizontal axis), according to its difficulty measure in the MAX-MIN sample.

Figure 5.2, which was generated using the Bond&FoxSteps3 software, shows the expected and empirical Item Characteristic Curves (ICCs) for 'engage 1', also based on the MAX-MIN sample. In this figure, the red line represents the expected ICC, while the blue line connecting 'x' markers shows the empirical curve. The expected ICC (the red curve indicated in **Figure 5.2**), corresponds to the logistic curve of the Rasch model. Hence, in the case of a dichotomous Rasch model, this curve provides a visual representation of equation (1). In the figure, the vertical axis indicates the probability of 'correct' response to a given item (P_{ni}), while the horizontal axis shows the measure of person ability relative to that item's difficulty ($\theta_n - \delta_i$). While the expected ICC is the same for every item –since they are characterized by the same (dichotomous) response structure–, the empirical ICC may vary, as it is based on the observed frequency of scores obtained on that specific item for each measure of ability. The boundary lines in the figure establish a (vertically interpreted) 95 % confidence band around the model curve. Points outside this band in the empirical ICC could suggest an unaccounted source of variance [162].

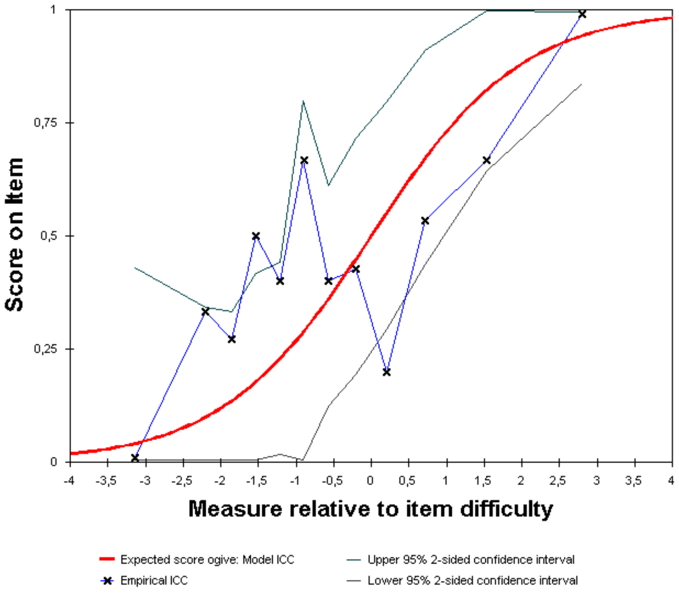

Figure 5.2. Joint display of the Expected and Empirical Item Characteristic Curves (ICC) for 'engage 1' in the MAX-MIN sample, indicating the probability of correct responses (vertical axis) given the difference of rater ability relative to the item's difficulty (horizontal axis), based on the Rasch model –the Expected ICC (represented by the red curve)– or on the observed data –the Empirical ICC (the blue line connecting 'x' markers). The boundary

lines around the Expected ICC show a (vertically interpreted) 95 % confidence band for the expected observations [162].

According to **Figure 5.2**, the responses significantly deviating from the expectations of the model in 'engage 1' were found for coders with ability levels a bit less than 0.5 logit above the item's difficulty (fixed at 0.0 logits), with their performance on that particular item being worse than expected, and for coders whose ability levels were around 1.5 logit below the item's difficulty, who presented better performance than expected on that item. Complementarily, as suggested by the results of the DIF analysis, coders from group A –who were particularly numerous at ability levels of around 1.5 logit below the difficulty measure for 'engage 1' (**Figure 5.1**)– performed better than expected on this item, while coders from group D –most of whom had ability levels a bit less than 0.5 logit above the item's difficulty (**Figure 5.1**)– performed worse.

For the highest construct level vignettes, six items presented misfit issues, with 'conslt 5' and 'countr 5' showing local dependence and overfit behaviour. The most difficult items, 'reason 3' and 'common 4', exhibited underfit and dimensionality problems, while indicators from V-Dem's Egalitarian principle, 'encmps 4' and 'univl 5', also showed local dependence and dimensionality issues.

Multidimensionality in the context of this study indicates that some items reveal an additional attribute of the coder. Based on the results from the principal component analysis of the Rasch model residuals (PCAR), one could suggest that recognising the highest levels of the constructs assessed by the selected group of indicators required different abilities than identifying the constructs at their lowest levels.

On the other hand, by limiting the sample to maximum and minimum level vignettes and converting the polytomous ratings into dichotomous scores based on their alignment with each vignette's construct level, lower ratings of minimum level vignettes were recorded as 1 (correct) and higher ratings, as zero (incorrect), whereas, for maximum level vignettes, the reverse was true: lower ratings were recorded as zero, and higher ratings as 1. This design might have led to an opposite trend between the responses to minimum and maximum level indicator-vignette pairs, with the dichotomous scores in the maximum level group following the same direction as the measurement of the construct assessed by the indicators (with 0 to

1 representing lower to higher construct levels), and the dichotomous scores in the minimum level group being inversely related to the indicators' latent variable (with 0 to 1 representing higher to lower construct levels).

The analysis of the item residuals in the first contrast shows some local dependence between item responses for the joint MAX-MIN sample, particularly associated with indicator vignettes at the highest level. The behaviour observed for the items "*Range of consultation 'conslt 5'*" and "*Respect counterarguments 'countr 5'*" is likely due to their overfit behaviour. A slight correlation was observed for the highest-level item-vignette pairs from the Egalitarian Principle, specifically "*Particularistic or public goods 'encmps 4'*" and "*Means-tested v. universalistic policy 'univl 5'*". These pairs also have the highest residual loadings in the 1st contrast, indicating the potential need for a common element to identify these constructs at their highest levels.

When considering the samples of responses to maximum and minimum level vignettes in isolation (MAX and MIN), deviating response patterns were associated with the most difficult indicator-vignette pairs –'engage 1' in the MIN sample and, less significantly, 'reason 3' and 'common 4' in the MAX sample.

The disattenuated correlation between clusters outside the limits of acceptance (clusters 1 and 3) and the one inside (cluster 2) raises concerns only for the Minimum Sample for Cluster 1 ('engage 1'). This item is the most difficult and underfit item-vignette pair in the survey and is associated with differential item functioning for raters from continents A and D. However, the residual loadings in the PCA first contrast suggest better responses for unidimensionality for the results provided by the MIN sample, with the worst presented by the MAX sample as shown in **Table 4.8**. This result is likely caused by the fact that most of the other item pairing low construct vignettes have better fitting behaviours, even though four occupy Rasch measure levels higher than zero compared to only two from the high construct vignettes.

Considering the dimensionality issues indicated in **Table 4.8** for the MAX sample and its residual cluster distribution joining the most challenging and most underfit items ('reason 3' and 'common 4'), these two items are considered the priorities for revision, followed by 'engage 1'.

The difficulty associated with 'engage 1', 'reason 3' and 'common 4' may be attributed to lapses in the coders' discrimination between specific rating categories, as is suggested by the frequency with which ratings were assigned to each indicator-vignette combination, shown in **Table 5.1**.

Table 5.1. Number of responses associating each of the fourteen indicator-vignette pairs to each rating category of the corresponding indicator, in each sample of respondents (MAX-MIN, MIN and MAX).

	sample of respondents	***rating categories:***						**Total number of respondents**
		0	**1**	**2**	**3**	**4**	**5**	
reason 1	MAX-MIN	19	**91**	22	10	-	-	142
	MIN	22	**94**	23	10	-	-	149
	MAX	19	**91**	23	10	-	-	143
reason 3	MAX-MIN	14	**48**	**64**	16	-	-	142
	MIN	14	**51**	**64**	17	-	-	146
	MAX	14	**48**	**65**	16	-	-	143
common 1	MAX-MIN	8	**82**	17	**25**	10	-	142
	MIN	8	**87**	17	**26**	11	-	149
	MAX	8	**82**	18	**25**	10	-	143
common 4	MAX-MIN	9	14	**26**	**57**	**36**	-	142
	MIN	9	15	**28**	**57**	**38**	-	147
	MAX	9	14	**26**	**58**	**36**	-	143
countr 1	MAX-MIN	18	**93**	17	3	8	3	142
	MIN	18	**99**	17	3	9	3	149
	MAX	18	**93**	17	4	8	3	143
countr 5	MAX-MIN	5	10	14	9	**26**	**78**	142
	MIN	5	11	14	9	**26**	**84**	149
	MAX	5	10	14	10	**26**	**78**	143
conslt 1	MAX-MIN	**46**	**77**	7	2	5	5	142
	MIN	**48**	**81**	7	2	5	6	149
	MAX	**46**	**77**	7	3	5	5	143
conslt 5	MAX-MIN	6	14	18	2	**70**	**32**	142
	MIN	6	14	19	2	**73**	**34**	148
	MAX	6	14	18	3	**70**	**32**	143
engage 1	MAX-MIN	14	**58**	**49**	6	10	5	142
	MIN	15	**60**	**53**	6	10	5	149
	MAX	14	**58**	**49**	6	11	5	143
engage 5	MAX-MIN	4	12	9	17	**41**	**59**	142
	MIN	4	12	9	18	**44**	**61**	148
	MAX	4	12	9	18	**41**	**59**	143
encmps 1	MAX-MIN	**27**	**72**	16	18	9	-	142
	MIN	**29**	**76**	16	18	10	-	149
	MAX	**27**	**72**	16	19	9	-	143
encmps 4	MAX-MIN	8	9	10	**44**	**71**	-	142
	MIN	8	10	10	**46**	**75**	-	149
	MAX	8	9	10	**44**	**72**	-	143
univl 1	MAX-MIN	**23**	**60**	**26**	7	18	8	142
	MIN	**24**	**63**	**28**	8	18	8	149
	MAX	**23**	**60**	**26**	7	18	8	142
univl 5	MAX-MIN	7	12	7	8	**87**	**21**	142
	MIN	7	12	9	9	**88**	**22**	147
	MAX	7	12	7	8	**87**	**22**	143

The difficulty encountered in recognizing the vignette depicting lowest level of the ‘engaged society’ indicator can be attributed to coders' lapses in discriminating between rating categories 1 and 2. While most respondents assigned the rating of 1 for the hypothetical case described in this vignette, category 2 was, by a narrow margin, the second most frequently selected option for this indicator-vignette pair.

The confusion between rating categories 1 and 2 of the 'engaged society' indicator could be associated with a difficulty in detecting nuances in the text of the vignettes[7] or of the questionnaire item. When addressing “how wide and how independent are public deliberations” [24] (**Table 4.2**), the response categories 1 and 2 of this indicator describe similar situations where public deliberation is allowed, but non-elite actors are typically kept out of the debate. The key difference between the two categories would then resume to the first two words in the description of category one, relating to the scope of public deliberation that is allowed [24] (p. 171):

> “1: *Some limited* ***public deliberations*** *are allowed but the public below the elite levels is almost always either unaware of major policy debates or unable to take part in them.*
>
> 2: ***Public deliberation*** *is not repressed but nevertheless infrequent and non-elite actors are typically controlled and/or constrained by the elites*.”

In other words, the difficulties in detecting the lowest level of the 'engaged society' indicator could have been influenced by difficulties in detecting —in the vignette’s text or in the description of the questionnaire item— the difference between restricted and unrepressed public deliberation.

As observed for the most challenging item ('engage 1'), the difficulty associated with 'reason 3' and 'common 4' may also be attributed to lapses in the coders' discrimination between specific rating categories. In the indicator for 'common good', two rating categories are remarkably similar [24] (pp. 169-170),

[7] The present study did not examine the vignettes' descriptions since their textual content was unavailable to the authors.

seemingly comprising, in both cases, situations in which justifications refer to “specific interests” almost as much as to the “common good”:

> *“2: Justifications are for the most part a **mix** of **specific interests** and the **common good** and it is impossible to say which justification is more common than the other.*
>
> *3: Justifications are based on a **mixture** of references to **constituency/party/group interests** and on appeals to the **common good**.”*

The confusion in the wording of the questionnaire item for the 'common good' indicator could have had a significant impact over coders' performance in recognizing the highest construct level vignette ('common 4', hypothetically located between the two highest rating categories, 3 and 4).

In the case of the indicator on 'reasoned justification', in turn, the significant difficulty in recognizing the highest construct level vignette ('reason 3'), but not the lowest ('reason 1'), could be related to difficulties in discriminating the quality of justification, as defined along the rating categories’ description for this particular item [24] (p. 169):

> *“1: **Inferior justification.** Elites tend to give reasons why someone should or should not be for doing or not doing something, but the reasons tend to be **illogical** or **false**, although they may appeal to many voters. For example, 'We must cut spending. The state is inefficient.' [The inference is incomplete because addressing inefficiencies would not necessarily reduce spending and it might undermine essential services.]*
>
> *2: **Qualified justification.** Elites tend to offer a **single simple reason** justifying why the proposed policies contribute to or detract from an outcome. For example, 'We must cut spending because taxpayers cannot afford to pay for current programs.'*
>
> *3: **Sophisticated justification.** Elites tend to offer more than one or more complex, nuanced and **complete** justification. For example, 'We must cut spending because taxpayers cannot afford to pay for current government programs. Raising taxes would hurt economic growth, and deficit spending would lead to inflation.'”*

In fact, when rating the vignette set at the highest level of 'reasoned justification' ('reason 3', hypothetically located between the categories 2 and 3), coders often marked it in level 1 (‘inferior justification’), almost as much as in the level 2 (‘qualified justification’) –as shown in **Table 4.2**. The difference between a justification that offers a “single simple reason” (“inferior justification”, described in Category 1) and one in which the “inference is incomplete” (“qualified justification”, in Category 2) may not be so clear; and identifying the reasons that are “illogical are false” may likewise involve a great deal of bias.

The significant difficulty in detecting the highest level of 'reasoned justification' ('reason 3') but not its lowest ('reason 1') –**Figure 4.2** and **Table 4.6**– suggest that coders towards more rigour when discriminating the quality of justifications in political discourses.

Figure 5.3 and **Figure 5.4** depict a graphical comparison between the distribution of coders' responses (listed in **Table 5.1**) for both 'reason 1' and 'reason 3' indicator-vignette pairs (**Figure 5.3**) and 'engage 1' and ' engage 5' (**Figure 5.4**).

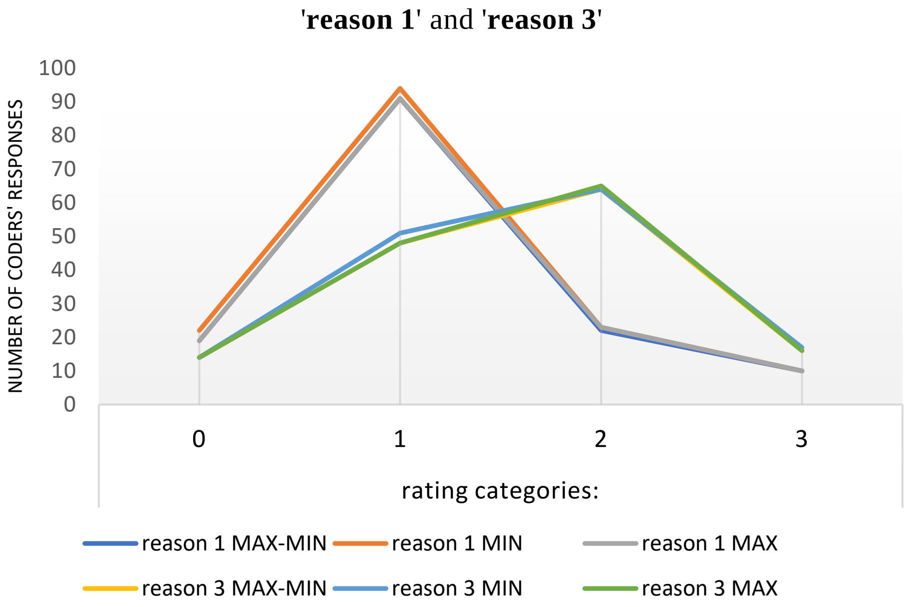

Figure 5.3. Number of coders' responses assigning the indicator-vignette pairs 'reason 1' and 'reason 3' to each rating category of the corresponding indicator, in each sample of respondents (MAX-MIN, MIN, and MAX)

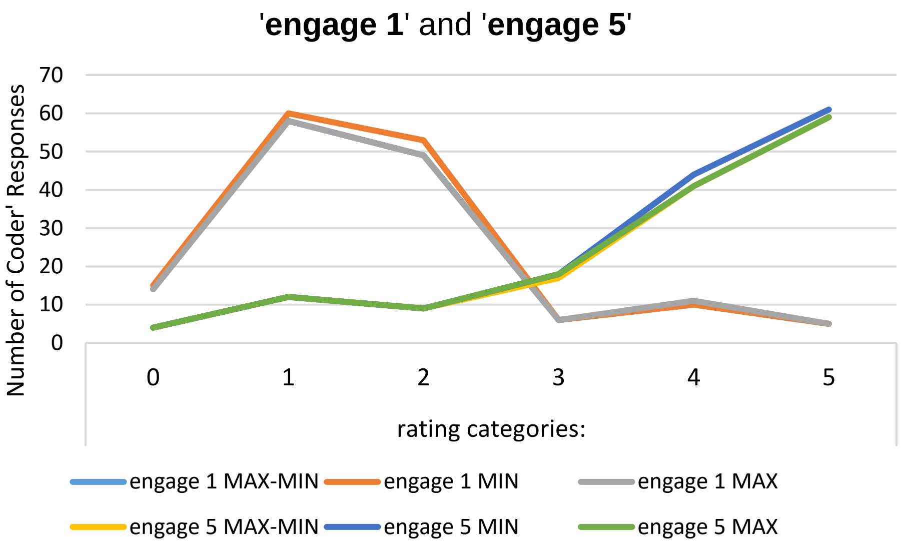

Figure 5.4. Number of coders' responses assigning the indicator-vignette pairs 'engage 1' and 'engage 5' to each rating category of the corresponding indicator, in each sample of respondents (MAX-MIN, MIN, and MAX)

For the 'reason 3' indicator-vignette pair (**Figure 5.3**), the responses are mainly divided between Category 2, which agrees with the vignette level provided, and Category 1, erroneously identified. In contrast, for 'reason 1', the pick of responses is concentrated on Category 1, which appropriately corresponds to the vignette construct level presented to the coders.

In turn, for the 'engage 1' indicator-vignette pair (**Figure 5.4**), the responses are mainly divided between Category 1, which agrees to the vignette level provided, and Category 2, erroneously identified, while for 'engage 5', the responses are distributed mainly in Categories 4 and 5, both agreeing to the highest vignette level being analysed.

Corroborating these results showing challenging items for respondents, the study described in [149], which investigates political trust surveys using Rasch, raised concerns about how the differences between the four answer categories were meaningful to the respondents. The robustness of the polytomous findings was then checked by dichotomising the items between categories 2 and 3. This simplification was shown not to affect the conclusions [149].

In the present study, the blurred discrimination between item step categories was generally observed in all indicator-vignette pairs, except for the item associated with ‘Respect Counterarguments’ (**'countr 1**' and '**countr 5**'), as shown in **Figure 5.5**.

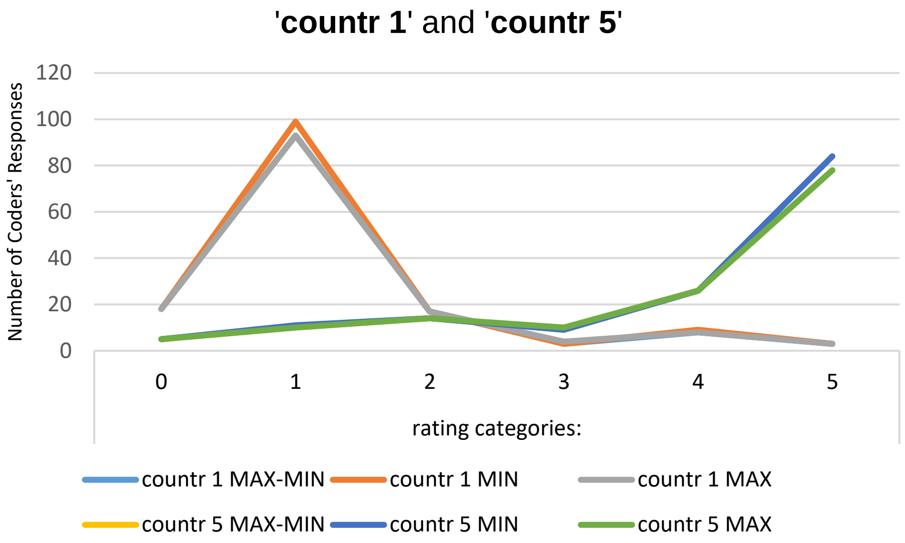

Figure 5.5. Number of coders' responses assigning the indicator-vignette pairs 'countr 1' and 'countr 5' to each rating category of the corresponding indicator, in each sample of respondents (MAX-MIN, MIN, and MAX)

For the item-vignette pair 'countr 1' (**Figure 5.5**), the pick of responses was correctly positioned in Category 1 without spreading significantly to the neighbour Category 2, which does not agree with the vignette level 1. In turn, for 'countr 5', raters massively and appropriately concentrated their responses in Category 5.

The inhomogeneity of response behaviours when exposed to the vignette text stimuli, the construct level's natural context, the wording of the questions, and the item response options provides crucial information for further investigation. This information includes the measurand, the potential unknown dimensions that can be identified, the specificities of individual culture and idiom groups, and the clearness of items and response options that can impact the raters' ability to identify the construct under measurement.

5.1 Conclusions and Future work

This preliminary investigation focused on the "sensor elements" of the democracy measuring system performed by the raters. These components were isolated for metrological characterisation by applying the Rasch probabilistic approach and analysing its parameters considering the complete set of indicators from the Deliberation Survey from the V-Dem measuring system and their anchoring vignettes database.

The study revealed various challenges that must be addressed to improve the measuring system's performance, ensuring the reproducibility and comparability of results. Given the complexity of the concept and its impact on item development, it is essential to continuously improve the clarity and specificity of indicator definitions. The study highlights specific items requiring investigation to improve text interpretation across coders, address differential functioning, and explore other interference parameters.

The analysis conducted using the Rasch approach allows for sample independence and comparability with future results provided by different groups of raters, as well as tracking their evaluations over time.

Overall, the results emphasise the intricate nature of assessing democracy and the importance of addressing multiple sources of uncertainty to enhance the accuracy of measurement results and provide measurement objectivity and intersubjectivity. The outcomes illustrate the impact of the sources of uncertainty in the measurement process on the performance of the measuring instrument's sensors. This understanding is crucial in shaping the infrastructure necessary to ensure the reliability of results from measuring systems designed for democracy assessments.

Efforts to integrate metrological considerations and strategies in this field of application can contribute to advancing research and theoretical discussions in political science.

Future work involves:

- Expanding the data analysis with anchoring vignettes.
- Incorporating the complete set of items and coders from the V-Dem project.
- Exploring the database with other approaches, notably the Many-Facet Rasch Model.
- Applying the Rasch probabilistic approach to the conventional measurement model of the V-Dem measuring system

References

[1] JCGM 200:2012, **International Vocabulary of Metrology – Basic and general concepts and associated terms (VIM)**, 3rd ed., Paris: Joint Committee for Guides in Metrology, 2012.

[2] WEBER, M. **Sobre la Teoría de las Ciencias Sociales**. Translation by Michael Faber-Kaiser. Barcelona: Península, v. 73, 1971.

[3] PLÜMPER, T.; NEUMAYER, E. Model uncertainty and robustness tests: Towards a new logic of statistical inference. **SSRN Electronic Journal**, 2012.

[4] BOND, T. G.; FOX, C. M. **Applying the Rasch Model**. [s.l.] Psychology Press, 2013.

[5] PENDRILL, L. Man as a measurement instrument. **NCSLi Measure**, v. 9, n. 4, p. 24-35, 2014.

[6] MARI, L.; WILSON, M. An introduction to the Rasch measurement approach for metrologists. **Measurement**, v. 51, p. 315-327, 2014.

[7] COSTA MONTEIRO, E. Bridging the boundaries between sciences to overcome measurement challenges. **Measurement: Interdisciplinary Research and Perspectives**, v. 15, n. 1, p. 34-36, 2017.

[8] MAUL, A.; MARI, L.; WILSON, M. Intersubjectivity of measurement across the sciences. **Measurement**, v. 131, p. 764-770, 2019.

[9] PENDRILL, L. **Quality assured measurement:** Unification across social and physical sciences. Springer, 2020.

[10] SALZBERGER, T.; CANO, S.; ABETZ-WEBB, L.; AFOLALU, E.; CHREA, C.; WEITKUNAT, R.; ROSE, J. Addressing traceability of self-reported dependence measurement through the use of crosswalks**,** **Measurement**, vol. 181, nº. 109593, 2021.

[11] GERRING, John; PEMSTEIN, Daniel; SKAANING, Svend-Erik. An ordinal, concept-driven approach to measurement: The lexical scale. **Sociological Methods & Research**, v. 50, n. 2, p. 778-811, 2021.

[12] FISHER, W. P.; CANO, S. J. (Ed.). **Person-centered outcome metrology:** Principles and applications for high stakes decision making. Springer, 2022.

[13] MARI, Luca; WILSON, Mark; MAUL, Andrew. Measurement across the sciences: Developing a shared concept system for measurement. Springer Nature, 2023.

[14] MONTEIRO VIEIRA, C.; COSTA MONTEIRO, E. Metrology in the early days of Social Sciences. **Acta IMEKO**, vol. 12, nº. 2, 2023, pp. 1-6.

[15] MONTEIRO VIEIRA, C.; COSTA MONTEIRO, E. Democracy measurement and metrology. In: **METROLOGIA 2023**, Petrópolis, Brazil, 28–30 Nov. 2023.

[16] MONTEIRO, Elisabeth Costa; SUMMERS, Ron. Metrological requirements for biomedical device assessment and their ethical implications. Measurement: Sensors, v. 24, p. 100574, 2022.

[17] POLLOCK III, Philip H.; EDWARDS, Barry C. **The essentials of political analysis**. Cq Press, 2019.

[18] DUNN, John. **Setting the people free:** The story of democracy. Princeton University Press, 2018.

[19] SKAANING, Svend-Erik; GERRING, John; BARTUSEVIČIUS, Henrikas. A lexical index of electoral democracy. **Comparative Political Studies**, v. 48, n. 12, p. 1491-1525, 2015.

[20] COPPEDGE, M.; GERRING, J.; LINDBERG, S. I.; SKAANING, S. E.; TEORELL, J. V-Dem comparisons and contrasts with other measurement projects. **V-Dem working paper** 45, 2017.

[21] COPPEDGE, M. et al. V-Dem methodology v14. **V-Dem Project**, 2024.

[22] MUNCK, Gerardo L.; VERKUILEN, Jay. Conceptualizing and measuring democracy: Evaluating alternative indices. **Comparative political studies**, v. 35, n. 1, p. 5-34, 2002.

[23] GRÜNDLER, Klaus; KRIEGER, Tommy. Using Machine Learning for measuring democracy: A practitioners guide and a new updated dataset for 186 countries from 1919 to 2019. **European Journal of Political**

Economy, v. 70, p. 102047, 2021.

[24] COPPEDGE, M. et al. V-Dem codebook v14. **V-Dem Project**, 2024.

[25] GIEBLER, Heiko; RUTH, Saskia P.; TANNEBERG, Dag. Why choice matters: revisiting and comparing measures of democracy. **Politics and Governance**, v. 6, n. 1, p. 1-10, 2018.

[26] VANHANEN, Tatu. A new dataset for measuring democracy, 1810-1998. **Journal of peace research**, v. 37, n. 2, p. 251-265, 2000.

[27] BÜHLMANN, Marc; MERKEL, Wolfgang; WESSELS, Bernhard. The quality of democracy: democracy barometer for established democracies. 2008.

[28] PEMSTEIN, Daniel; MESERVE, Stephen A.; MELTON, James. Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type. **Political Analysis**, v. 18, n. 4, 2010.

[29] CHEIBUB, José Antonio; GANDHI, Jennifer; VREELAND, James Raymond. Democracy and dictatorship revisited. **Public choice**, v. 143, p. 67-101, 2010.

[30] BÜHLMANN, Marc et al. The democracy barometer: A new instrument to measure the quality of democracy and its potential for comparative research. **European Political Science**, v. 11, p. 519-536, 2012.

[31] BOIX, Carles; MILLER, Michael; ROSATO, Sebastian. A complete data set of political regimes, 1800–2007. **Comparative political studies**, v. 46, n. 12, p. 1523-1554, 2013.

[32] BJØRNSKOV, Christian; RODE, Martin. Regime types and regime change: A new dataset on democracy, coups, and political institutions. **The Review of International Organizations**, v. 15, n. 2, p. 531-551, 2020.

[33] ENGLER, Sarah et al. **Democracy Barometer codebook - version 7**. 2020.

[34] House Freedom 2020 methodology Freedom in the World

[35] Marshall, M 2020 Polity5: Dataset Users' Manual (Center of Systematic Peace)

[36] BOESE, Vanessa A. How (not) to measure democracy. **International Area Studies Review**, v. 22, n. 2, p. 95-127, 2019.

[37] MARQUARDT, Kyle L.; PEMSTEIN, Daniel. IRT models for expert-coded panel data. **Political Analysis**, v. 26, n. 4, p. 431-456, 2018.

[38] BAKKER, Ryan et al. The European common space: Extending the use of anchoring vignettes. **The Journal of Politics**, v. 76, n. 4, p. 1089-1101, 2014.

[39] BOLLEN, Kenneth A.; PAXTON, Pamela. Subjective measures of liberal democracy. **Comparative political studies**, v. 33, n. 1, p. 58-86, 2000.

[40] CASPER, Gretchen; TUFIS, Claudiu. Correlation versus interchangeability: The limited robustness of empirical findings on democracy using highly correlated data sets. **Political Analysis**, v. 11, n. 2, p. 196-203, 2003.

[41] SEAWRIGHT, Jason; COLLIER, David. Rival strategies of validation: Tools for evaluating measures of democracy. **Comparative Political Studies**, v. 47, n. 1, p. 111-138, 2014.

[42] MUNCK, Gerardo L. What is democracy? A reconceptualization of the quality of democracy. **Democratization**, v. 23, n. 1, p. 1-26, 2016.

[43] FISHMAN, Robert M. Rethinking dimensions of democracy for empirical analysis: Authenticity, quality, depth, and consolidation. **Annual Review of Political Science**, v. 19, n. 1, p. 289-309, 2016.

[44] MARI, Luca. Beyond the representational viewpoint: a new formalization of measurement. **Measurement**, v. 27, n. 2, p. 71-84, 2000.

[45] WILSON, Mark. Using the concept of a measurement system to characterize measurement models used in psychometrics. **Measurement**, v. 46, n. 9, p. 3766-3774, 2013.

[46] PENDRILL, Leslie; PETERSSON, Niclas. Metrology of human-based and other qualitative measurements. **Measurement Science and Technology**, v. 27, n. 9, p. 094003, 2016.

[47] CANO, S. J. et al. Patient-centred cognition metrology. **Journal of Physics: Conference Series**, v. 1065, n. 7, p. 072033, 2018.

[48] PENDRILL, L. **Quality assured measurement**. Springer International Publishing, 2019.

[49] FISHER JR, William P.; MASSENGILL, Paula J. **Explanatory**

models, unit standards, and personalized learning in educational measurement: Selected papers by A. Jackson Stenner. Springer Nature, 2023.

[50] BIPM, The International System of Units, 9th ed., Sèvres: International Bureau of Weights and Measures, 2019. Online [Accessed 10 July 2022] https://www.bipm.org/en/publications/si-brochure

[51] COSTA MONTEIRO, E.; LEON, L. F. Metrological reliability of medical devices. **Journal of Physics: Conference Series**, v. 588, n. 1, p. 012032, 2015. DOI: https://doi.org/10.1088/1742-6596/588/1/012032

[52] COSTA MONTEIRO, Elisabeth. Bridging the boundaries between sciences to overcome measurement challenges. **Measurement: Interdisciplinary Research and Perspectives**, v. 15, n. 1, p. 34-36, 2017. DOI: https://doi.org/10.1080/15366367.2017.1358974

[53] COSTA MONTEIRO, Elisabeth. Magnetic quantities: healthcare sector measuring demands and international infrastructure for providing metrological traceability. **TMQ – Techniques, Methodologies and Quality**, pp. 42-50, 2019.

[54] BRISTOW, Adrian F. Assignment of quantities to biological medicines: an old problem re-discovered. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 369, n. 1953, p. 4004-4013, 2011. DOI: https://doi.org/10.1098/rsta.2011.0175

[55] DAMATTA, R. **Relativizando:** Uma introdução a antropologia social. Rocco, Rio de Janeiro, 2010, ISBN 8532501540.

[56] GUTTMAN, Louis. The basis for scalogram analysis. **Scaling**, p. 142, 1974.

[57] GUTTMAN, Louis. A basis for scaling qualitative data. **American sociological review**, v. 9, n. 2, 1944, pp. 139-150.

[58] GUTTMAN, Louis. An approach for quantifying paired comparisons and rank order. **The Annals of Mathematical Statistics**, v. 17, n. 2, pp. 144-163, 1946.

[59] FISHER JR, William P.; PENDRILL, Leslie (Ed.). **Models, Measurement, and Metrology Extending the SI:** Trust and Quality Assured Knowledge Infrastructures. Walter de Gruyter GmbH & Co KG, 2024.

[60] MARI, Luca; UGAZIO, Erica. Preliminary analysis of validation of measurement in soft systems. Journal of Physics: Conference Series, v. 238, n. 1, p. 012026, 2010. DOI: https://doi.org/10.1088/1742-6596/238/1/012026

[61] FISHER JR, William P.; STENNER, A. Jackson. Metrology for the social, behavioral, and economic sciences. Social, Behavioral, and Economic Sciences White Paper Series). Online [Accessed 10 July 2022] http://www.truevaluemetrics.org/DBpdfs/Metrics/William-P-Fisher/FisherJr_William_Metrology-for-the-Social-Behavioral-and-Economic-Sciences.pdf

[62] MARI, Luca; CARBONE, Paolo; PETRI, Dario. Fundamentals of hard and soft measurement. In: **Modern Measurements:** Fundamentals and Applications. A. Ferrero, D. Petri, P. Carbone, M. Catelani (Ed.). Wiley-IEEE Press, 2015, pp. 203-262.

[63] DJURIC, Mladen; FILIPOVIC, Jovan; KOMAZEC, Stefan. Reshaping the future of social metrology: Utilizing quality indicators to develop complexity-based scientific human and social capital measurement model. **Social Indicators Research**, v. 148, 2020, pp. 535-567. DOI: https://doi.org/10.1007/s11205-019-02217-6

[64] DELMASTRO, Marco. **On the Measurement of Social Phenomena:** A Methodological Approach. Springer International Publishing, 2021.

[65] FISHER JR, W. P. Almost the Tarde model? **Rasch Measurement Transactions**, v. 28, n. 1, 2014, pp. 1459-1461. Online [Accessed 10 July 2022] https://www.rasch.org/rmt/rmt281.pdf

[66] FISHER JR, W. P. The central theoretical problem of the social sciences. **Rasch Measurement Transactions**, v. 28, n. 2, 2014, pp. 1464-1466. Online [Accessed 10 July 2022] http://www.rasch.org/rmt/rmt282.pdf

[67] FULLMER, Susanna; DANIEL, David. **History and Development of Psychometrics**. 2020.

[68] ENGELHARD JR, George. **Invariant measurement:** Using Rasch models in the social, behavioral, and health sciences. Routledge, 2013.

[69] KÆRGÅRD, N. Georg Rasch and modern econometrics. In: **Seventh Scandinavian History of Economic Thought Meeting**, Molde University College, Molde, Norway. 2003.

[70] FISHER JR, William P. Invariance and traceability for measures of human, social, and natural capital: Theory and application. **Measurement**, v. 42, n. 9, p. 1278-1287, 2009. DOI: https://doi.org/10.1016/j.measurement.2009.03.014

[71] ZHONG, Hua; XU, Jianhua; PIQUERO, Alex R. Internal migration, social exclusion, and victimization: An analysis of Chinese rural-to-urban migrants. **Journal of research in crime and delinquency**, v. 54, n. 4, p. 479-514, 2017. DOI: https://doi.org/10.1177/0022427816676861

[72] MELIN, Jeanette et al. Construct specification equations:'Recipes' for certified reference materials in cognitive measurement. **Measurement: Sensors**, v. 18, p. 100290, 2021. DOI: https://doi.org/10.1016/j.measen.2021.100290

[73] DA ROCHA, Neusa Sica et al. An introduction to Rasch analysis for psychiatric practice and research. **Journal of psychiatric research**, v. 47, n. 2, p. 141-148, 2013. DOI: https://doi.org/10.1016/j.jpsychires.2012.09.014

[74] UHER, Jana. Measurement in metrology, psychology and social sciences: data generation traceability and numerical traceability as basic methodological principles applicable across sciences. **Quality & Quantity**, v. 54, n. 3, p. 975-1004, 2020. DOI: https://doi.org/10.1007/s11135-020-00970-2

[75] WRIGHT, Benjamin D. A history of social science measurement. Educational measurement: **Issues and practice**, v. 16, n. 4, p. 33-45,

1997. DOI: https://doi.org/10.1111/j.1745-3992.1997.tb00606.x

[76] FISHER JR, William P.; STENNER, A. Jackson. Theory-based metrological traceability in education: A reading measurement network. **Measurement**, v. 92, 2016, pp. 489-496. DOI: https://doi.org/10.1016/j.measurement.2016.06.036

[77] BAIRD, Jo-Anne et al. Metrology of education. Assessment in Education: Principles, **Policy & Practice**, v. 24, n. 3, 2017, pp. 463-470. DOI: https://doi.org/10.1080/0969594X.2017.1337628

[78] RASCH, G**. Probabilistic Models for some Intelligence and Attainment Tests**. University of Chicago Press, Chicago, 1980. ISBN 978-0226705538.

[79] RASCH, Georg. On general laws and the meaning of measurement in psychology. In: **Proceedings of the fourth Berkeley symposium on mathematical statistics and probability**, 1961, pp. 321-333.

[80] MARI, Luca. Is our understanding of measurement evolving. **Acta IMEKO**, v. 10, n. 4, 2021, pp. 209-213. DOI: http://dx.doi.org/10.21014/acta_imeko.v10i4.1169

[81] WILSON, Mark; FISHER, William. Joint IMEKO TC1-TC7-TC13 Symposium: Metrology Across the Sciences: Wishful Thinking? **Journal of Physics: Conference Series**. v. 772, 2016, p. 011001. DOI: https://doi.org/10.1088/1742-6596/772/1/011001

[82] COSTA MONTEIRO, E. 2017 Joint IMEKO TC1-TC7-TC13 Symposium: Measurement Science Challenges in Natural and Social Sciences. **Journal of Physics: Conference Series**, v. 1044, 2018, p. 011001. DOI: https://doi.org/10.1088/1742-6596/1044/1/011001

[83] WILSON, Mark; FISHER, William. Preface of the special issue, Psychometric Metrology. **Measurement**, v. 145, 2019, p. 190-190. DOI: 10.1016/j.measurement.2019.05.077

[84] STEVENS, Stanley Smith. On the theory of scales of measurement. Science, v. 103, n. 2684, 1946, p. 677-680.

[85] MONTEIRO VIEIRA, Clara; COSTA MONTEIRO, Elisabeth. (no

prelo). Rasch model and anchoring vignettes for metrological characterization of a democracy measuring system: preliminary studies. In: **XXIV IMEKO World Congress "Think Metrology"**, Hamburg, Alemanha, 26–29 ago. 2024

[86] ALPERT, Harry. **Emile Durkheim and his sociology**. Columbia University Press, 1939. ISBN 9780231909983.

[87] DURKHEIM, Émile. **Les règles de la méthode sociologique**. Revue Philosophique de la France et de l'Étranger, v. 37, 1894, p. 465-498.

[88] DURKHEIM, Émile. **De la division du travail social** (1893). Presses Universitaires France, 2007, ISBN 978-2130563297.

[89] DELLA PORTA, Donatella; KEATING, Michael (Ed.). Approaches and methodologies in the social sciences: A pluralist perspective. Cambridge University Press, 2008. ISBN 978-0521709668.

[90] DURKHEIM, E. **Le Suicide: Étude de Sociologie** (1897). Hachette Livre Bnf, 2013. ISBN 978-2012895508.

[91] TARDE, Gabriel. **Essais et mélanges sociologiques**. Paris: Maloine, 1895, pp. 196-197.

[92] TARDE, Gabriel. **La croyance et le désir: la possibilité de leur mesure**. 1876.

[93] ANTER, Andreas; BREUER, Stefan. **Max Webers Staatssoziologie:** Positionen und Perspektiven. Nomos Verlagsgesellschaft, 2007. ISBN 978-3832927738.

[94] WEBER, Max. **Methodology of social sciences** (1903-1917). Routledge, 2017. ISBN 978-1138528048.

[95] LLANQUE, M. Max Weber, wirtschaft und gesellschaft. Grundriss der verstehenden soziologie, Tübingen 1922, in: Schlüsselwerke der Politikwissenschaft. S. Kailitz (editor). VS Verlag für Sozialwissenschaften, 2007, ISBN 978-3-531-90400-9, pp. 489-493. [In German]

[96] HOLTON, Robert. Max Weber and the interpretative tradition. In: Handbook of Historical Sociology, 2003, pp. 27-38. G. Delanty, E. F.

Isin (editors). SAGE, London, 2003, ISBN 978-0761971733, pp. 27-38.

[97] WEBER, Max. **The Protestant Ethic and the Spirit of Capitalism** [1904–5]. Merchant Books, 2013. ISBN 9781603866040.

[98] COSER, Lewis A. **Masters of Sociological Thought:** Ideas in Historical and Social Context, 2nd ed. Harcourt Brace Jovanovich, New York, 1977. ISBN 0155551302 9780155551305.

[99] ADCOCK, Robert; COLLIER, David. Measurement validity: A shared standard for qualitative and quantitative research. **American political science review**, v. 95, n. 3, 2001, pp. 529-546

[100] BOLLEN, Kenneth A. **Structural equations with latent variables**. New York: John Wiley, 2014.

[101] ANDRICH, David. Controversy and the Rasch model: a characteristic of incompatible paradigms? **Medical care**, v. 42, n. 1, 2004, pp. I7-I16.

[102] MCGRANE, Joshua A.; MAUL, Andrew. The human sciences, models and metrological mythology. **Measurement**, v. 152, 2020, p. 107346

[103] ARYADOUST, Vahid; NG, Li Ying; SAYAMA, Hiroki. A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. **Language Testing**, v. 38, n. 1, 2021, pp. 6-40.

[104] TUTZ, Gerhard. Invariance of comparisons: Separation of item and person parameters beyond Rasch models. **Journal of Mathematical Psychology**, v. 122, 2024, pp. 102876.

[105] ANDRICH, David; MARAIS, Ida. **A course in Rasch measurement theory:** Measuring in the educational, social and health sciences, v. 41, 2019.

[106] TESIO, Luigi et al. Interpreting results from Rasch analysis 2: Advanced model applications and the data-model fit assessment. **Disability and rehabilitation**, v. 46, n. 3, 2024, pp. 604-617.

[107] HAGELL, Peter. Testing rating scale unidimensionality using the principal component analysis (PCA)/t-test protocol with the Rasch model: the primacy of theory over statistics. **Open Journal of Statistics**, v. 4, n. 6, 2014, pp. 456-465.

[108] TESIO, Luigi et al. Interpreting results from Rasch analysis 1: The "most likely" measures coming from the model. **Disability and rehabilitation**, v. 46, n. 3, 2024, pp. 591-603.

[109] LINACRE, John M.; WRIGHT, Benjamin D. The "length" of a logit. **Rasch Measurement Transactions**, v. 3, n. 2, 1989, pp. 54-55.

[110] LORD, F. Fred Lord and Ben Wright discuss Rasch and IRT Models. **Rasch Measurement Transactions**, v. 24, n. 3, 2010, pp. 1289-1290.

[111] WRIGHT, Benjamin D.; MOK, Magdalena MC. An overview of the family of Rasch measurement models. **Introduction to Rasch measurement**, v. 1, n. 1, p. 1-24, 2004, pp. 1-24.

[112] WRIGHT, B. D.; MASTERS, G. N. **Rating scale analysis**. Chicago: MESA Press, 1982.

[113] MASTERS, G. N. A Rasch model for partial credit scoring. **Psychometrika**, v. 47, 1982, pp. 149-174.

[114] LINACRE, J. M. **Many-facet Rasch measurement**. Chicago: MESA Press, 1989.

[115] LINDBERG, Staffan I. et al. V-Dem: A new way to measure democracy. **Journal of Democracy**, v. 25, n. 3, 2014, pp. 159-169

[116] TEORELL, Jan et al. Measuring electoral democracy with V-Dem data: Introducing a new polyarchy index. **V-Dem Working Paper** 25, 2016.

[117] COPPEDGE, Michael et al. Measuring high level democratic principles using the V-Dem data. **International Political Science Review**, v. 37, n. 5, 2016, pp. 580-593

[118] GLEDITSCH, Kristian S.; WARD, Michael D. Double take: A reexamination of democracy and autocracy in modern polities. **Journal of Conflict Resolution**, v. 41, n. 3, 1997, pp. 361–383

[119] ELFF, M.; ZIAJA, S. Method Factors in Democracy Indicators. **Politics and Governance**, v. 6, n. 1, 2018, pp. 92–104

[120] ELKINS, Zachary. Gradations of democracy? Empirical tests of alternative conceptualizations. **American Journal of Political Science**, 2000, pp. 287-294

[121] VACCARO, Andrea. Comparing measures of democracy: statistical properties, convergence, and interchangeability. **European Political Science**, v. 20, n. 4, 2021, pp. 666-684

[122] LANDMAN, Todd. Democracy and human rights: Concepts, measures, and relationships. **Politics and Governance**, v. 6, n. 1, 2018, pp. 48-59.

[123] KOELBLE, Thomas A.; LIPUMA, Edward. Democratizing democracy: A postcolonial critique of conventional approaches to the ‘measurement of democracy’. **Democratisation**, v. 15, n. 1, 2008, pp. 1-28.

[124] GRUGEL, Jean. Democratization studies: citizenship, globalization and governance. **Government and Opposition**, v. 38, n. 2, 2003, pp. 238-264.

[125] BROOKS, Heidi; NGWANE, Trevor; RUNCIMAN, Carin. Decolonising and re-theorising the meaning of democracy: A South African perspective. **The Sociological Review**, v. 68, n. 1, p. 17-32, 2020, pp. 17-32.

[126] FLEUß, Dannica; HELBIG, Karoline; SCHAAL, Gary S. Four parameters for measuring democratic deliberation: Theoretical and methodological challenges and how to respond. **Politics and Governance**, v. 6, n. 1, 2018, pp. 11-21.

[127] FUCHS, Dieter; ROLLER, Edeltraud. Conceptualizing and measuring the quality of democracy: The citizens' perspective. **Politics and Governance**, v. 6, n. 1, 2018, pp. 22–32.

[128] MAYNE, Quinton; GEIßEL, Brigitte. Don’t good democracies need “good” citizens? Citizen dispositions and the study of democratic quality. **Politics and Governance**, v. 6, n. 1, 2018, pp. 33-47.

[129] COPPEDGE, Michael; REINICKE, Wolfgang H. Measuring polyarchy. **Studies in Comparative International Development**, v. 25, 1990, pp. 51-72

[130] SKAANING, Svend-Erik. Different types of data and the validity of democracy measures. **Politics and Governance**, v. 6, n. 1, 2018, pp. 105–116.

[131] COPPEDGE, Michael; ALVAREZ, Angel; MALDONADO, Claudia. Two persistent dimensions of democracy: Contestation and inclusiveness. **The Journal of Politics**, v. 70, n. 3, 2008, pp. 632–647.

[132] LAUTH, Hans-Joachim. **The matrix of democracy:** a three-

dimensional approach to measuring the quality of democracy and regime transformations. Würzburg: Universität Würzburg, 2015.

[133] COPPEDGE, M.; REINICKE, W. H. On measuring democracy: Its consequences and concomitants. In: A. Inkeles (ed.). **On measuring democracy: Its consequences and concomitants**. New Brunswick: Transaction, 1991, pp 47-68.

[134] BOLLEN, Kenneth. Liberal democracy: Validity and method factors in cross-national measures. **American Journal of Political Science**, 1993, pp. 1207-1230.

[135] SHEN, Ce; WILLIAMSON, John B. Corruption, democracy, economic freedom, and state strength: A cross-national analysis. **International journal of comparative sociology**, v. 46, n. 4, 2005, pp. 327-345

[136] TREIER, Shawn; JACKMAN, Simon. Democracy as a latent variable. **American Journal of Political Science**, v. 52, n. 1, 2008, pp. 201-217.

[137] PRZEWORSKI, Adam. **Democracy and Development:** Political Institutions and Well-Being in the World, 1950-1990. Cambridge University Press, 2000.

[138] YASHAR, Deborah J. **Demanding democracy:** Reform and reaction in Costa Rica and Guatemala, 1870s-1950s. Stanford University Press, 1997.

[139] MAHONEY, James. **The Legacies of Liberalism:** Path Dependence and Political Regimes in Central America. Baltimore: Johns Hopkins University Press, 2001.

[140] COLLIER, David. Data, field work, and extracting new ideas at close range. **Newsletter of the Organized Section in Comparative Politics of the American Political Science Association**, v. 10, n. 1, 1999, pp. 4-6.

[141] RUESCHEMEYER, Dietrich; STEPHENS, Evelyne Huber; STEPHENS, John D. **Capitalist Development and Democracy**. University Of Chicago press. 1992.

[142] BOWMAN, Kirk; LEHOUCQ, Fabrice; MAHONEY, James. Measuring political democracy: Case expertise, data adequacy, and Central America. **Comparative Political Studies**, v. 38, n. 8, 2005, pp. 939-970.

[143] O'DONNELL, Guillermo A. Illusions about consolidation. **Journal of democracy**, v. 7, n. 2, 1996, pp. 34-51.

[144] COLLIER, David; LEVITSKY, Steven. Democracy with adjectives: Conceptual innovation in comparative research. World politics, v. 49, n. 3, 1997, pp. 430-451.

[145] MAINWARING, Scott; BRINKS, Daniel; PÉREZ-LIÑÁN, Aníbal. Classifying political regimes in Latin. **Studies in Comparative International Development**, v. 36, 2001, pp. 37-65.

[146] BAKER, Peter J.; KOESEL, Karrie J. Measuring “polyarchy plus”: Tracking the quality of democratization in Eastern Europe. In: **annual meeting of the American Political Science Association**, San Francisco, CA. 2001.

[147] MØLLER, Jørgen; SKAANING, Svend-Erik. Beyond the radial delusion: Conceptualizing and measuring democracy and non-democracy. **International Political Science Review**, v. 31, n. 3, 2010, pp. 261-283.

[148] BOLLEN, Kenneth A. Issues in the comparative measurement of political democracy. **American Sociological Review**, 1980, pp. 370–390.

[149] VAN DER MEER, Tom WG; OUATTARA, Ebe. Putting ‘political’ back in political trust: an IRT test of the unidimensionality and cross-national equivalence of political trust measures. **Quality & quantity**, v. 53, n. 6, 2019, pp. 2983-3002.

[150] DAHL, Robert A. **Polyarchy: Participation and Opposition**. Yale University Press: New Haven, 1971.

[151] DAHL, Robert A. **Democracy and its Critics**. Yale University Press: New Haven, 1989.

[152] MARQUARDT, Kyle L. et al. Experts, Coders, and Crowds: An analysis of substitutability. **V-Dem Working Paper** 53, 2017.

[153] LINACRE, John M. **A user's guide to WINSTEPS® MINISTEP:** Rasch-model computer programs. Program Manual 3.68. 0. Chicago, IL, 2009.

[154] LINACRE, John M. Data variance explained by Rasch measures. **Rasch Measurement Transactions**, v. 20, n. 1, 2006, pp. 1045.

[155] LINACRE, J. M. Variance in data explained by Rasch measures. **Rasch Measurement Transactions**, v. 22, n. 3, p. 1164, 2008, p. 1164. [http://www.rasch.org/rmt/rmt221j.htm].

[156] BOONE, William J.; STAVER, John R. **Advances in Rasch analyses in the human sciences**. Cham, Switzerland: Springer, 2020.

[157] CHOU, Yeh-Tai; WANG, Wen-Chung. Checking dimensionality in item response models with principal component analysis on standardized residuals. **Educational and Psychological Measurement**, v. 70, n. 5, 2010, pp. 717-731.

[158] MARI, Luca; CARBONE, Paolo; PETRI, Dario. Measurement fundamentals: a pragmatic view. **IEEE transactions on instrumentation and measurement**, v. 61, n. 8, 2012, pp. 2107-2115.

[159] RAÎCHE, Gilles. Critical eigenvalue sizes in standardized residual principal components analysis. **Rasch measurement transactions**, v. 19, n. 1, 2005, p. 1012.

[160] TABATABAEE-YAZDI, Mona et al. Development and Validation of a Teacher Success Questionnaire Using the Rasch Model. **International Journal of Instruction**, v. 11, n. 2, 2018, p. 129-144.

[161] J. A. Shaffer, D. Degeest, L. I. Andrew, Tackling the problem of construct proliferation: A guide to assessing the discriminant validity of conceptually related constructs. Organizational Research Methods, vol. 19, nº. 1, 2016, pp. 80-110.

[162] Linacre, J. M. **Expected+Empirical ICC or IRF**. Online [Accessed 22 September 2024] https://winsteps.com/facetman/expectedempirical_icc.htm