**ValiTex Checklist**

**Use Case E: Zero-Shot/Few-Shot Classification (unknown output categories)**

This checklist accompanies the [ValiTex](https://arxiv.org/abs/2307.02863) framework for validating text-based measures of social science constructs by Birkenmaier et al. (2023).   
Each row within the table corresponds to one validation step (i.e., a single reported and clearly demarcated validation activity). Validation steps can be either **recommended** or **optional** depending on their relevance. As outlined in the corresponding paper, researchers should initially follow the order of the phases, starting with the substantive validation steps and ending with external validation steps while continuously considering robustness checks. However, researchers might adapt this process to their individual use case.

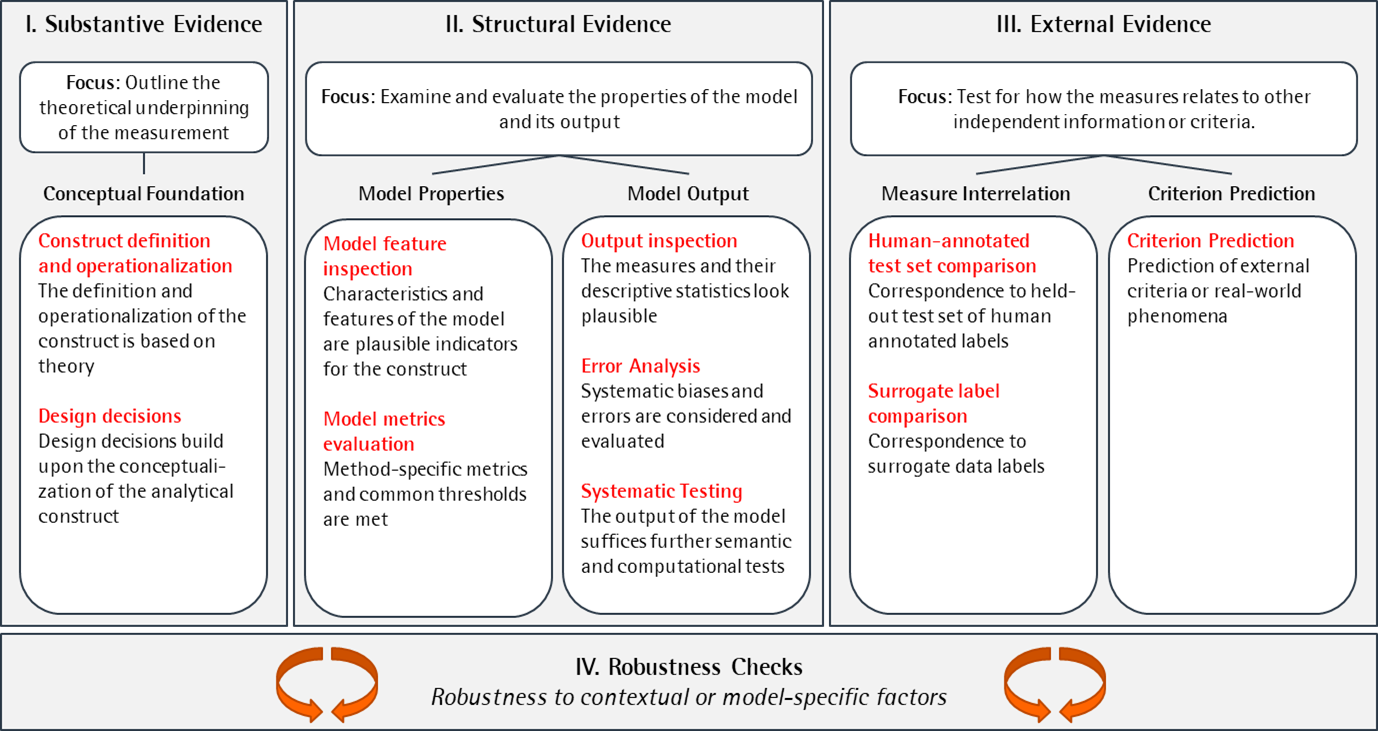


Figure 1: Framework

# Substantive Evidence

Before conducting any measurements, researchers need to outline the theoretical under-pinning of the measurement to demonstrate substantive evidence. Validation steps for **substantive evidence** should therefore demonstrate that the measurement is based on a strong conceptual foundation, including the operationalization of the construct and the design decisions around the measurement process.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Validation Step** | **Documentation** | **Status** | **Considerations** | **Performance Criteria** | **Source / References** |
| I.1 | Documentation of the conceptual background |  | R | Have I conducted a literature review or consulted with domain experts to gain a comprehensive understanding of conceptual background the construct (e.g., dimensionality)? | Evidence of engagement with the construct | Krippendorf (2018), Chapter 9 (Analytical Construct);  Clark and Watson (2019) |
| I.2 | Justification of the operatio-nalization |  | R | Have I sufficiently explained how the construct should manifest itself in the textual data? | Theoretical reasoning | Krippendorf (2018), Chapter 9 (Analytical Construct) |
| I.3 | Manual Precoding |  | O | Have I conducted a pilot study using manual coding to evaluate the inter-rater agreement and reliability on detecting the construct by hand? | Agreement between coders | Krippendorf (2018), Chapter 11 (Reliability) |
| I.4 | Justification of data collection decisions |  | R | Have I selected a dataset that is representative and relevant to the research question and population of interest? Have I justified the data selection decisions (e.g., using keywords)? Have I assessed the quality and completeness of the dataset and checked for potential biases or inconsistencies? | Theoretical reasoning | Grimmer et al. (2022), Chapter 4 (Selecting Documents); Sen et al. (2021) |
| I.5 | Justification of method choice |  | R | Have I selected the appropriate type of method based on the operationalization of the construct and data characteristics? Have I justified the concrete selection of a particular model, and have I documented relevant features of the model? | Theoretical reasoning | Grimmer & Steward (2013); Grimmer et al. (2022) |
| I.6 | Justification of the level of analysis |  | R | Have I selected the appropriate level of analysis? Have I considered potential problems when aggregating scores from lower to higher levels (e.g., sentence to paragraph level)? | Theoretical reasoning | For an application, see Jankowski & Huber (2022) |
| I.7 | Justification of preprocessing decisions |  | R | Have I justified the preprocessing decisions, such as removing stopwords, based on the presumed manifestation of the construct in the text? | Theoretical reasoning | Grimmer et al. (2022), Chapter 5 (Bag of Words);  Denny and Spirling (2018) |

# Structural Evidence

For **structural evidence**, researchers should conduct validation steps to examine and evaluate the properties of the model and its output. Structural evidence enables the researcher to gain a deeper understanding of how the measurement process functions, as well as to identify any biases or errors introduced by the computational workflow.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Validation Step** | **Documentation** | **Status** | **Considerations** | **Performance Criteria** | **Source / References** |
| II.1 | Inspection of predictive features |  | R | Have I considered interpretable machine learning techniques such as LIME, ICE, or partial dependence? Have I considered the interpretability and relevance of the top-ranked features? | Subjective assessment | R |
| II.4 | Inspection of classifiers confidence strengths |  | O | Have I assessed the model's confidence level on individual predictions? | Subjective assessment | O |
| II.7 | Visual inspection of output |  | R | Have I visualized my output? Have I identified and visualized outliers and extreme values? | Subjective assessment | R |
| II.8 | Visual inspection of measures over time |  | O | Have I plotted the temporal trends of my measures and assessed their stability and consistency over time? | Subjective assessment | O |
| II.9 | Comparison of aggregated measures across known groups |  | R | Have I compared the aggregated measures across known groups (e.g., across data characteristics or subsets of the data)? | Subjective assessment | R |
| II.10 | Comparison of data features for clusters of closely related measures |  | O | Have I compared important data features, such as the average length of text or how often certain words appear together, across texts with similar scores (e.g., same classes on a discrete scale or high/low values on a continuous scale)? | Subjective assessment | O |
| II.11 | Reading top documents with the highest overall scores for each output category |  | R | Have I read the most outstanding documents for each type of output, such as such as for distinct groups or topics, or highest and lowest scores on a numerical scale? | Subjective assessment | R |
| II.12 | Conducting error analysis using data grouping |  | R | Have I conducted an error analysis to compare the performance of my model across known subgroups? | Subjective assessment | R |
| II.13 | Conducting qualitative error analysis of outstanding or deliberatively chosen observations |  | R | Have I conducted an error analysis to qualitatively evaluate the sources and types of errors associated with the measures? | Subjective assessment | R |
| II.14 | Conducting functional tests |  | O | Have I designed and conducted functional tests (i.e., manually prepared text samples) to evaluate the model's ability to detect specific patterns in a realistic or simulated scenario? | Metric assessment | O |
| II.15 | Conducting adversarial or counterfactual tests |  | O | Have I designed and conducted adversarial or counterfactual tests to ensure that my model is sensitive to changes in the text data? | Metric assessment | O |
| II.16 | Conducting computational text intrusion tasks |  | O | Have I designed and conducted computational intrusion tasks to whether the model is able to recognize texts unrelated to the construct of interest (e.g., by assigning low scores on the construct)? | Metric assessment | O |

# External Evidence

For external evidence, researchers should conduct validation steps that test for how the measures corresponds to independent information or criteria. Thus, information outside the scope of the textual data in which the measure was constructed serves as an external benchmark (hence “external” evidence).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Validation Step** | **Documentation** | **Status** | **Considerations** | **Performance Criteria** | **Source / References** |
| III.1 | Comparison of measures with human-annotated test set ("gold-standard data") |  | R | Have I labelled a subset of the data using a codebook or pairwise comparison method to serve as the gold standard for evaluation? | Correspondence to human-annotated test set | Grimmer et al. (2022), Chapter 20 (Checking Performance); Song et al. (2020) |
| III.2 | Comparison of measures with surrogate labels |  | O | Have I collected or generated surrogate labels (e.g., expert surveys, contextual labels) as another benchmark for evaluation? | Correspondence to surrogate labels | Adcock and Collier (2013);  Grimmer et al. (2022), Chapter 20 (Checking Performance) |
| III.3 | Prediction of external criteria or real-world phenomena |  | O | Have I formulated expected relationship of my measures with external criteria? Have I confirmed these relationships empirically? | Correspondence to external criteria | Adcock and Collier (2013);  Grimmer et al. (2022), Chapter 20 (Checking Performance) |

# Robustness Checks

Next to the three types of validation evidence outlined above, the ValiTex framework also recommends the continuous test of robustness checks to assess the impact of researchers’ degree of freedom on the measurement outcome. On a general note, one could see robustness checks as additional means to test whether decisions regarding the measurement design might have a sustainable effect on the measure’s outcome.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Validation Step** | **Documentation** | **Status** | **Considerations** | **Performance Criteria** | **Source / References** |
| IV.1 | Rerunning the analysis using different preprocessing steps |  | R | Have I rerun the analysis using different preprocessing settings (e.g., stop word removal, stemming, lemmatization)? | Change to previous measurement outcome | Denny and Spirling (2018) |
| IV.2 | Rerunning the analysis using different hyperparameter settings |  | R | Have I rerun the analysis using different hyperparameter settings? | Change to previous measurement outcome | Arnold et al. (2023) |
| IV.3 | Rerunning the analysis using alternative text-based methods |  | O | Have I rerun the analysis with alternative text-based methods? | Change to previous measurement outcome | For an application, see van Atteveldt et al. (2021) |
| IV.4 | Rerunning the analysis with different levels of aggregation |  | O | Have I replicated the same study using different levels of aggregation (e.g., token, word, sentence, paragraph, document level)? | Change to previous measurement outcome | For an application, see Boukes et al. (2019) |
| IV.5 | Rerunning the analysis with a different, but related dataset |  | O | Have I replicated the same study using a different, but related dataset? | Change to previous measurement outcome | Grimmer et al. (2022), Chapter 20 (Checking Performance) |
| IV.6 | Rerunning the analysis using different subsets of the data |  | O | Have I rerun the analysis using different subsets of the data? | Change to previous measurement outcome | For an application see Yarchi et al. (2020) |
| IV.7 | Rerunning the analysis using different thresholds |  | O | Have I rerun the analysis using different thresholds (e.g., min. number of tokens matched, max. document frequency)? | Change to previous measurement outcome | For an application, see Baden et al. (2020) |

# Literature

Arnold, C., Biedebach, L., Küpfer, A., & Neunhoeffer, M. (2023). The role of hyperparameters in machine learning models and how to tune them. *Political Science Research and Methods*.

Baden, C., Kligler-Vilenchik, N., & Yarchi, M. (2020). Hybrid content analysis: Toward a strategy for the theory-driven, computer-assisted classification of large text corpora. *Communication Methods and Measures*, *14*(3), 165–183.

Boukes, M., Velde, B., Araujo, T., & Vliegenthart, R. (2019). What’s the Tone? Easy Doesn’t Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools. *Communication Methods and Measures*, *14*, 1–22. https://doi.org/10.1080/19312458.2019.1671966

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological assessment*, *31*(12), 1412.

Denny, M. J., & Spirling, A. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis*, *26*(2), 168–189. https://doi.org/10.1017/pan.2017.44

Goet, N. D. (2019). Measuring polarization with text analysis: Evidence from the UK House of Commons, 1811–2015. *Political Analysis*, *27*(4), 518–539.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267–297.

Jankowski, M., & Huber, R. A. (2022). When Correlation Is Not Enough: Validating Populism Scores from Supervised Machine-Learning Models. *Political Analysis*, 1–15.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin’words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, *16*(4), 372–403.

Müller, S. (2020). Media Coverage of Campaign Promises Throughout the Electoral Cycle. *Political Communication*, *37*(5), 696–718. https://doi.org/10.1080/10584609.2020.1744779

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021). HateCheck: Functional Tests for Hate Speech Detection Models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 41–58. https://doi.org/10.18653/v1/2021.acl-long.4

Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., & Boomgaarden, H. G. (2020). In Validations We Trust? The Impact of Imperfect Human Annotations as a Gold Standard on the Quality of Validation of Automated Content Analysis. *Political Communication*, *37*(4), 550–572. https://doi.org/10.1080/10584609.2020.1723752

van Atteveldt, W., van der Velden, M. A., & Boukes, M. (2021). The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, *15*(2), 121–140.

Wu, T., Ribeiro, M. T., Heer, J., & Weld, D. (2019). Errudite: Scalable, Reproducible, and Testable Error Analysis. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 747–763. https://doi.org/10.18653/v1/P19-1073

Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2021). Political Polarization on the Digital Sphere: A Cross-platform, Over-time Analysis of Interactional, Positional, and Affective Polarization on Social Media. *Political Communication*, *38*(1–2), 98–139. https://doi.org/10.1080/10584609.2020.1785067

Zhang, J., Wang, Y., Molino, P., Li, L., & Ebert, D. S. (2018). Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE transactions on visualization and computer graphics*, *25*(1), 364–373.