**ValiTex Checklist**

**Method: Supervised Methods**

This checklist accompanies the [ValiTex](https://github.com/lukasbirki/ValiTex-Checklist) 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.

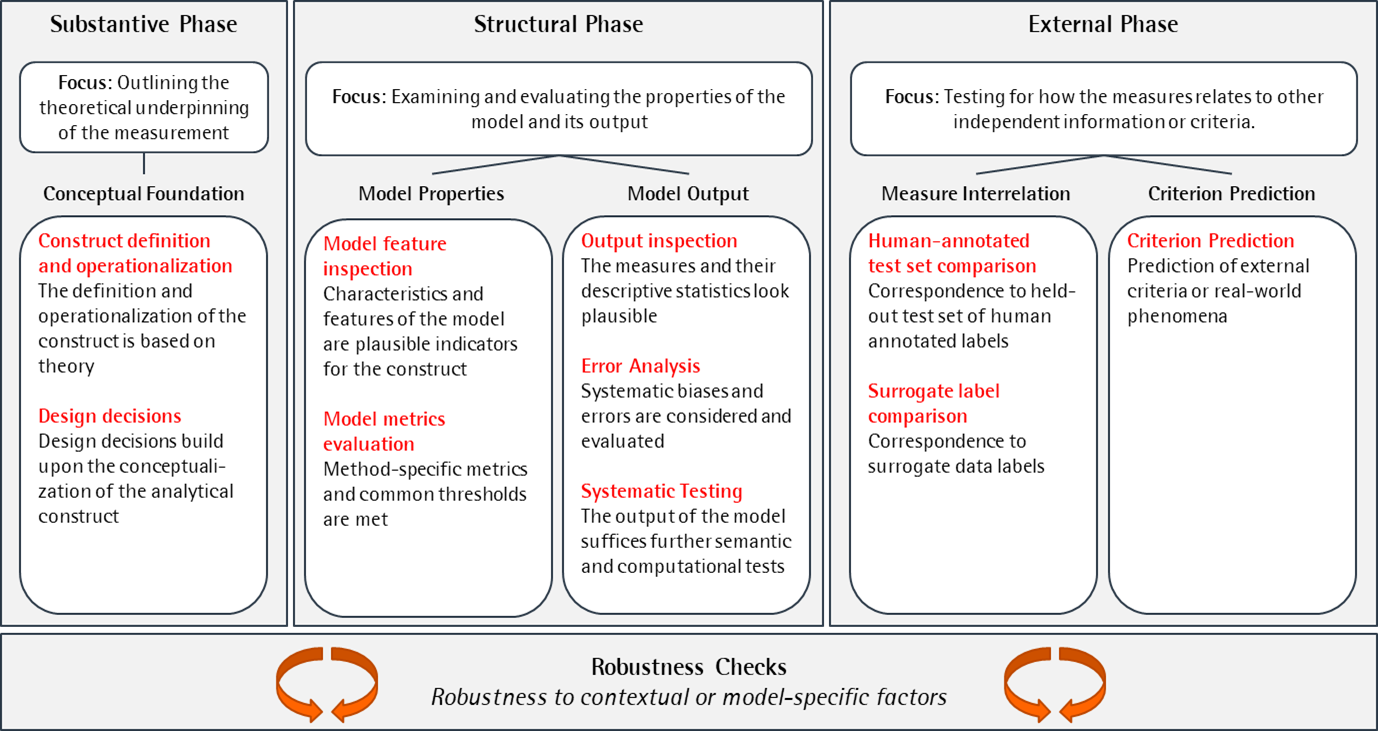


Figure 1: Conceptual Model

# Substantive Phase

The substantive phase involves outlining the theoretical underpinning of the measurement. Validation steps within the substantive phase 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 Phase

The structural phase involves validation steps to examine and evaluate the properties of the model and its output. The objective of the structural phase is 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 | Molnar (2020); Lindartos et al. (2020);  For a practical introduction, see Awesome XAI (2023) |
| II.6 | Inspection of classifiers confidence strengths |  | O | Have I assessed the model's confidence level on individual predictions? | Subjective assessment | Moradi & Samwald (2022); Andersen & Maalej (2022) |
| II.7 | Visual inspection of output |  | R | Have I visualized my output? Have I identified and visualized outliers and extreme values? | Subjective assessment | For an application, see Goet (2019) |
| 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 | For an application, see Goet (2019) |
| 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 | For an application, see Goet (2019) |
| II.10 | Comparison of data features for clusters of closely related measures |  | R | 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 | For an application, see Müller (2020) |
| 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 | For an application, see Goet (2019) |
| 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 | Zhang et al. (2018);  Wu et al. (2019) |
| 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 | Zhang et al. (2018);  Wu et al. (2019) |
| II.14 | Conducting functional tests |  | R | 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 | Röttger et al. (2021) |
| 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 |  |
| II.16 | Conducting computational text intrusion tasks |  | R | 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 | For an application, see Huang et al. (2020) |

# External Phase

In the external phase, researchers should ultimately test for how the measures relates to independent information or criteria.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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 phases of the validation process outlined above, one fundamental principle of the ValiTex framework also includes the continuous test of robustness checks to assess the impact of researchers’ degree of freedom. On a general note, one could see robustness checks as additional means to test whether design decisions in the substantive phase 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

Tba.