**Valitex Checklist Supervised Methods**

This checklist accompanies the [ValiTex](https://github.com/lukasbirki/ValiTex-Checklist" \t "_blank) framework for validating text-based measures of social science constructs by Birkenmaier et al. (2023 forthcoming).   
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.

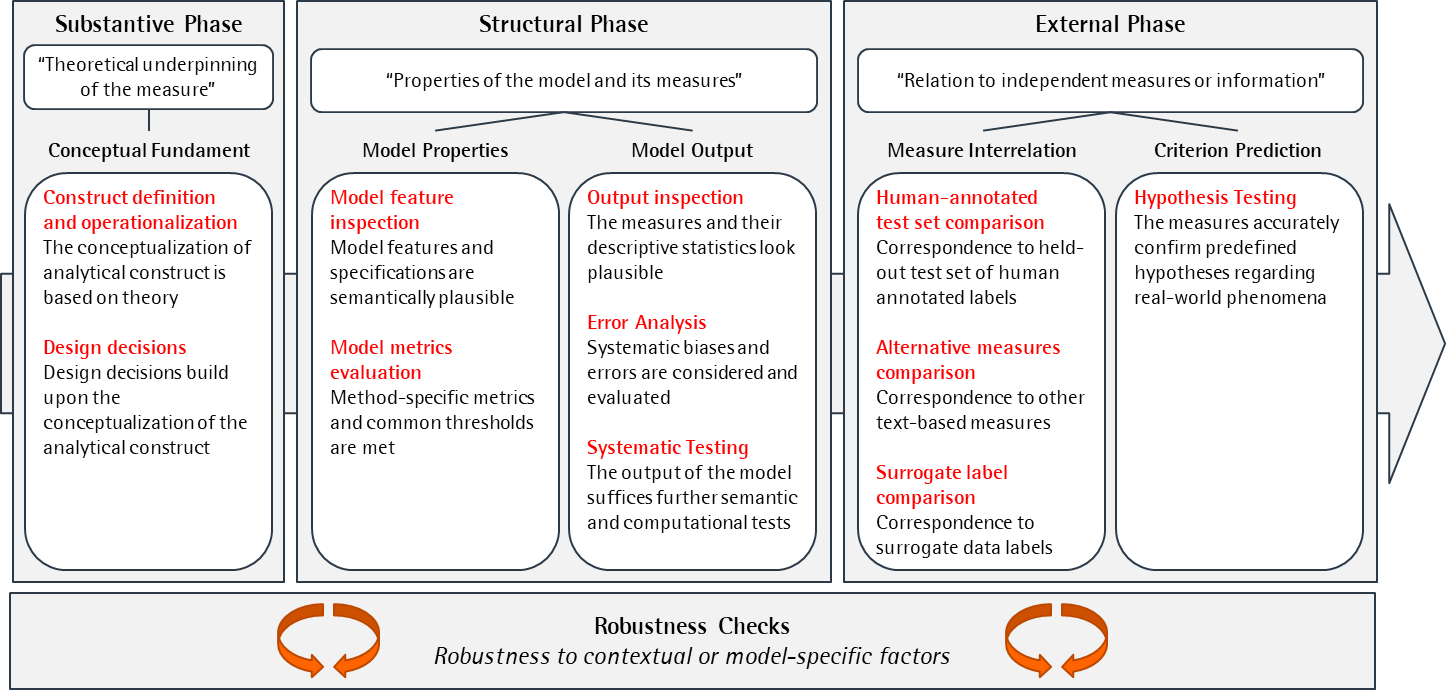


Figure 1: Conceptual Model

# Substantive Phase

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Validation Step** | **Documentation** | **Status** | **Considerations** | **Performance Criteria** | **Further Information / Literature** |
| Competence building (context, manifestation, dimensionality) |  | Recommended | Have I conducted a literature review or consulted with domain experts to gain a comprehensive understanding of the construct? | Evidence of engagement with the construct |  |
| Theoretical justification for the link between the construct and the textual data |  | Recommended | Have I justified my operationalization in regard to the greater context and the manifestation and dimensionality of my construct within the textual data? | Conceptual argumentation |  |
| Empirical justification for the link between the construct and the textual data using human precoding |  | Optional | 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 |  |
| Justification on the choice of method |  | Recommended | Have I selected the appropriate method based on the research question, data characteristics, and model assumptions? | Conceptual argumentation |  |
| Justification on data collection decisions (e.g., keywords identification) |  | Recommended | 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? Have I assessed the quality and completeness of the dataset and checked for potential biases or confounding variables? | Conceptual argumentation |  |
| Justification on the level of analysis (token, word, sentence, paragraph, or document level) |  | Recommended | Have I selected the appropriate level of analysis based on the research question, data size, and computational resources? Have I considered potential problems when aggregating scores from lower to higher levels (e.g., sentence to text level) | Conceptual argumentation |  |
| Justification on the choice of preprocessing decisions, such as removing parts of the data and data cleaning |  | Recommended | Have I justified the preprocessing decisions based on the research question, data quality, and/or computational efficiency? | Conceptual argumentation |  |

# Structural Phase

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Validation Step** | **Documentation** | **Status** | **Considerations** | **Performance Criteria** | **Further Information / Literature** |
| Conducting feature importance analysis |  | Recommended | Have I considered different feature importance metrics and their effects on the model's performance? Have I considered the interpretability and relevance of the top-ranked features in the context of the research question? | Subjective assessment |  |
| Inspecting XAI visualizations (LIME, ICE, Partial Dependence, Breakdown Plot) |  | Recommended | Have I chosen appropriate XAI techniques based on the model complexity, data size, and research question? Have I checked for the stability and consistency of the explanations across different subsets of data? | Subjective assessment |  |
| Inspecting confidence strengths of classifiers |  | Optional | Have I assessed the model's confidence level on individual predictions and analysed the distribution of confidence scores? | Subjective assessment |  |
| Visual inspection of outliers / measures with extreme values |  | Recommended | Have I identified and visualized the outliers or extreme values in the text data or analysis results? | Subjective assessment |  |
| Visual inspection of measures over time |  | Optional | Have I plotted the temporal trends of relevant measures (e.g., topic prevalence, sentiment score) and assessed their stability and consistency over time?   Have I identified the potential causal factors or events that may influence the trends and validated the relationships using statistical tests or domain knowledge? | Subjective assessment |  |
| Visual inspection of differences between groups (e.g., mean text length, top words used) |  | Recommended | Have I compared the relevant text measures (e.g. ideological positions or sentiment scores) between different groups and visualized the differences? | Subjective assessment |  |
| Conducting error analysis of outstanding or deliberatively chosen measures |  | Recommended | Have I conducted an error analysis to evaluate the sources and types of errors associated with the measures and identified potential improvements or remedies? | Subjective assessment |  |
| Conducting computational text intrusion tasks |  | Optional | Have I designed and conducted computational intrusion tasks to evaluate the model's sensitivity and specificity to various types of perturbations in the text data?   Have I analysed the intrusion results to identify the key features or patterns that the model relies on and validated their relevance and consistency with the research question and domain knowledge? | Metric assessment |  |
| Conducting functional tests for construct detection |  | Optional | Have I designed and conducted functional tests to evaluate the model's ability to detect specific patterns in a realistic or simulated scenario? | Metric assessment |  |

# External Phase

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Validation Step** | **Documentation** | **Status** | **Considerations** | **Performance Criteria** | **Further Information / Literature** |
| Comparing measures with human-annotated test set ("gold-standard data") |  | Recommended | Have I collected and labelled a subset of the data using a codebook or pairwise comparison method to serve as the gold standard for evaluation? | Agreement between coders |  |
| Comparing measures with alternative text-based methods |  | Optional | Have I compared the measures generated by the model with alternative text-based methods (e.g., topic modelling, sentiment analysis, word embeddings) to assess their interrelationships? | Correlation with alternative measure scores |  |
| Comparing measures with surrogate labels (e.g., expert surveys; contextual labels, etc.) |  | Optional | Have I collected or generated surrogate labels (e.g., expert surveys, contextual labels) to evaluate the model's performance and generalization to new data or domains? | Correspondence to surrogate labels |  |
| Testing of substantive and established hypotheses |  | Optional | Have I formulated substantive and established hypotheses about the relationship of my measures with external criteria? Have I confirmed these relationship empirically? | Empirical correspondence to external criteria |  |

# Robustness Checks

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Validation Step** | **Documentation** | **Status** | **Considerations** | **Performance Criteria** | **Further Information / Literature** |
| Rerunning the analysis using different preprocessing steps |  | Recommended | Have I rerun the analysis using different preprocessing settings (e.g., stop word removal, stemming, lemmatization) to evaluate their impact on the results? | Change to previous measurement outcome |  |
| Rerunning the analysis using different hyperparameter settings |  | Recommended | Have I rerun the analysis using different hyperparameter settings to evaluate their impact on the results? | Change to previous measurement outcome |  |
| Replication of the same study with different levels of aggregation |  | Optional | Have I replicated the same study using different levels of aggregation (e.g., token, word, sentence, paragraph, document level) to evaluate their impact on the results? | Change to previous measurement outcome |  |
| Replication of the same study with a different, but related dataset |  | Optional | Have I replicated the same study using a different, but related dataset to evaluate the generalizability and robustness of the analysis? | Change to previous measurement outcome |  |
| Rerunning the analysis using different subsets of the data |  | Optional | Have I rerun the analysis using different subsets of the data to evaluate the robustness and generalizability of the model? | Change to previous measurement outcome |  |
| Rerunning the analysis using different thresholds |  | Optional | Have I rerun the analysis using different thresholds (e.g., min. number of tokens matched, max. document frequency) to evaluate their impact on the results? | Change to previous measurement outcome |  |

# Literature

Tba.