Evaluating Evaluation Metrics: A Framework for Analyzing NLG Evaluation Metrics using Measurement Theory

Anonymous EMNLP submission

Abstract

We address a fundamental challenge in Natural Language Generation (NLG) model evaluation—the design and validation of evaluation metrics. Recognizing the limitations of existing automatic metrics and noises in how current human evaluation was conducted, we propose a new framework backed by measurement theory, the foundation of educational test design, for conceptualizing and evaluating the *reliability* and *validity* of NLG evaluation metrics. The framework formalizes the source of measurement error in metrics and offers statistical tools for assessing metrics based on empirical data. To exemplify the use of our framework in practice, we analyzed a set of evaluation metrics for summarization and identified the conflated validity structure in human-eval and reliability issues in LLM-based metrics. Through our framework, we aim to promote the design, evaluation, and interpretation of valid and reliable metrics to advance robust and effective NLG models. [1]

1 Introduction

For natural language generation (NLG) models, evaluation metrics provide quantitative assessments to guide development, benchmark scientific progresses, and inform generalizability across tasks and domains (Novikova et al., 2017). Effective evaluation metrics can extract valuable signals and robust evidence from model outputs that identify the strengths and weaknesses of different models, allowing for more informed decision-making in real-world deployment (Zhou et al., 2022). Conversely, problematic evaluation metrics can mislead development and deployment, resulting in downstream harms to individuals and society (Yeo and Chen, 2020; Sheng et al., 2021).

However, designing effective evaluation metrics for NLG tasks has long been challenging due to the complex nature of language, open-endedness of tasks, multifacet and context-dependent definition of language quality(Nema and Khapra, 2018; Zhou et al., 2022; Gehrmann et al., 2022; Sai et al., 2022). To cope with these challenges, NLG evaluation metrics have evolved from word-based metrics (e.g., ROUGE (Lin, 2004), BLEU (Papineni et al., 2002)) to embedding-based metrics (e.g., BERTScore (Zhang et al., 2019), MoverScore (Zhao et al., 2019)) and to end-to-end metrics (e.g., BLEURT (Sellam et al., 2020), G-Eval (Liu et al., 2023)). Most recently, the NLG evaluation challenge has been further exacerbated by the emergence of "general-purpose" large language models (LLMs), further demanding evaluation to capture model utility for diverse downstream use cases.

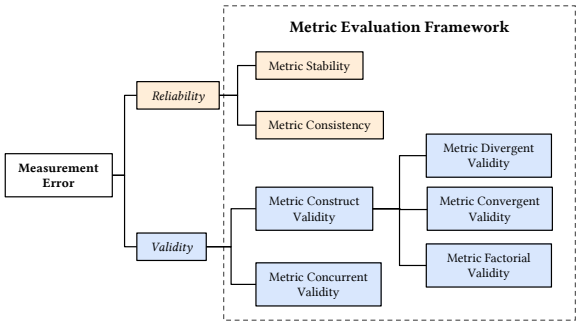

Figure 1: Metric Evaluation Framework. The framework conceptualized and operationalized four main components that evaluate two main sources of measurement error,*reliability* and *validity*, in measurement theory.

With this increasingly rich set of evaluation metrics being pursued, we must understand how *good* each of them is. While researchers pointed out shortcomings of popular metrics (e.g., ROUGE), such as their inability to capture semantic meanings, insensitivity to perturbations, and failure to reflect real-world performance (Sai et al., 2021; Liu et al., 2016; Reiter, 2018; Celikyilmaz et al., 2020; Kauchak and Barzilay, 2006), there is a lack of principled approaches to evaluate NLG evaluation

[1] Data and code for analysis are available at: anonymous.during.review

metrics, and to begin with, a lack of clear definition on what makes a metric good.

Some prior works have attempted to evaluate the quality of NLG evaluation metrics by their correlations with human judgments (Sai et al., 2021; Fabbri et al., 2021; Liu et al., 2016), which are deemed the gold standard for quality assessment. However, correlation with human preferences gives limited quality signals. More problematically, human evaluation data collection itself currently suffers from validation, standardization, consistency, and reproducibility issues (Clark et al., 2021; Howcroft et al., 2020; Belz et al., 2021; Khashabi et al., 2021). These issues subsequently undermine their validity as the foundation for evaluating automatic metrics.

In this paper, we introduce a *Metric Evaluation Framework* to define the desiderata of and assess evaluation metrics by drawing from measurement theory in educational and psychological testing. Based on two core concepts in measurement theory that define a "good" metric in testing individual capabilities—*reliability* and *validity*, our framework lays out four desiderata: Metric Stability, Metric Consistency, Metric Construct Validity, and Metric Concurrent Validity. We further propose a set of statistical tools to quantify these desiderata to systematically evaluate evaluation metrics. We demonstrate our framework with a case study of evaluating how 16 types of NLG metrics perform on a summarization task.

This paper offers three contributions,

- Introduce and transfer metrics evaluation desiderata and methods from measurement theory in educational and psychological testing to NLG evaluation.
- Propose a theory-driven framework with a set of statistical tools for systemically analyzing and evaluating NLG metrics.
- A case study demonstrating how to apply our framework and identify issues of evaluation metrics for a summarization task.

2 Measurement Theory

Originating from educational and psychological testing, measurement theory aims to inform evaluation processes that devise a coherent numerical representation of individual capabilities. For instance, a person's language proficiency (e.g., essay responses to questions). Scores on these tests have direct consequences for high-stakes decisions, such as school admissions.

Key to measurement theory is the distinction between the *observed score* on a test, e.g., the score of an examinee's essay in a language proficiency exam, and the *true score* on the general *construct* (Cronbach and Meehl, 1955) that the test is theorized to measure, e.g., language proficiency. The gap between the observed and true scores is referred to as *measurement error* (Allen and Yen, 2001). Measurement theory defines two sources of measurement error. Random measurement errors are fluctuations specific to a time, place, examinee, and assessment form that are transient and balance out to 0 over repeated measures, and they have direct consequences on the *reliability* of the evaluation process. Systematic errors, other the other hand, are persistent shifts across one or more time, place, examinee, or assessment form, and they have direct consequences on test *validity* by producing observed scores with systematic deviations from the true score that the test purports to measure (e.g., a downward bias if the rubric on a language proficiency exam looks for specialized knowledge about a certain subject).

By evaluating and identifying the source of measurement errors, the test designer could iteratively improve their test design by adding or removing test items or changing their rubric. The results can also help the evaluator to interpret a test score with caution. For example, for tests with lower reliability, the evaluator may administer the test multiple times and use their average score.

In short, as a safeguard to the trustworthiness of tests, measurement theory offers a conceptual framework for how the validity and reliability of a test should be formalized, evaluated, and optimized to reduce measurement error with the aid of statistical methods and tools.

2.1 Transferring Measurement Theory to the Context of NLG

There are obvious analogies between the measurement of human capability and the evaluation of NLG models. When evaluating a model, we similarly hope to derive scores based on the candidate's observed performance on a tailored collection of tasks, e.g., benchmark, so as to (1) draw inferences about the candidate's underlying ability in a specific domain (e.g., summarization), and (2) provide the evaluator guidance on the candidate's expected

behavior in future tasks of the domain. Similar to educational testing, the score is often interpreted and used beyond its nominal meaning, i.e., implying the model's general performance beyond the particular benchmark dataset.

The conceptual and statistical tools provided by measurement theory can be transferred to assist in evaluating NLG metrics, specifically to quantify and identify different sources of measurement errors. Not only can these tools help the community systematically assess the shortcomings of evaluation metrics and identify misleading ones, but they also guide the interpretation of their evaluation results, as well as the re-design of existing metrics and the development of new ones. In the next section, we elaborate on how we transfer these tools from measurement theory to a framework that defines and assesses the reliability and validity of NLG evaluation metrics, and how they may help us interpret and improve NLG evaluation.

3 Metric Evalaution Framework

In this section, we introduce each component of our framework that defines and assesses different aspects of the "goodness" of NLG evaluation metrics, inspired by the core concept of reliability and validity in measurement theory. For example, to evaluate a summarization model, one can apply reference-based metrics (e.g., ROUGE, BertScore), reference-free metrics (e.g., SUPERT), or human ratings on specific output quality aspects (e.g., coherence or relevance) on the same benchmark to draw inferences about model capability. Our framework is interested in evaluating and comparing the reliability and validity of the metrics. For the remainder of this section, we will illustrate our framework with this running example of evaluating summarization models with diverse metrics.

It is important to note that the quality of evaluation results is also dependent on the chosen dataset and reference (for reference-based metrics), which, in NLG evaluation, are concerned with benchmark designs. Measurement errors may cascade from those components to the observed score. In this work, we focus on the metrics part only and answer questions such as "giving a CNN/Daily Mail benchmark, does using ROUGE or BertScore or human ratings offer reliable and valid evaluation results". This is an important question given the far-reaching impact that prevalent benchmarks can have on the output of the NLP community.

3.1 Reliability

The reliability of a metric is the extent to which the result is subject to random measurement error and thus *(in)consistent* across repeated measures, such as different (sub-)datasets within a benchmark or different raters scoring the model's output in human evaluation. Suppose two NLG models are scored on their performance based on Metric-A on a summarization benchmark. Researchers and practitioners often use the scores to draw inferences about the models' (relative) performances. When the two models are reported to differ in their scores (e.g., Metric-A = .39 vs .42), a natural question is to what extent this reflects actual differences (true signal) versus fluctuations due to random measurement error (noise). If Metric-A is unreliable, the measurement error may mislead the comparison.

Sources of random measurement error that impair the reliability of a metric may include:

- Non-deterministic algorithms of some metrics may produce score variations on the same model outputs.
- The subsets of data points (e.g., different genres of articles) included in the benchmark.
- For human evaluations, the variability across raters, resulting from their subjectivity, inconsistency, errors, and so on;

In classical test theory (Spearman, 1904), the observed metric score of a model (X) is equal to the sum of true score T and error (E), which is assumed to be independent of T and fluctuates around 0 with variance σ_E^2. The goal of evaluating a metrics' reliability is hence to quantify the expected amount of fluctuation in the observed score due to random measurement error, known as the standard error of measurement (σ_E).

To empirically estimate σ_E is via the reliability coefficient of a metric, denoted $\rho_{XT}^2 \in [0, 1]$. Formally, the reliability coefficient is defined as the proportion of variance in the observed score explained by the variance in the true score across NLG models, or equivalently, the squared correlation between X and T:

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}. \quad (1)$$

Metrics with higher reliability coefficients are more desirable. However, in reality, neither T nor E is

observed. The reliability coefficient in Equ. 1 cannot be directly computed and is statistically approximated via several possible estimators.

Our framework proposes to estimate the reliability coefficient from two angles: Metric Stability and Metric Consistency. They reflect different reliability issues (whether a metric remains stable for the same model outputs v.s. whether it remains consistent with dataset perturbations) that can arise in different types of metrics, as we elaborate below. By quantifying and identifying reliability issues, metric developers can improve the scoring algorithms, and metric users can make more informed decisions in choosing metrics and adopt mitigation strategies, e.g., increasing the test set size to mitigate consistency issues (Spearman, 1910).

3.1.1 Metric Stability

Metric Stability refers to how a metric score may fluctuate when evaluated again on the same model output. While we would expect perfect stability (i.e., $\sigma_E = 0$) for deterministic metrics, such as ROUGE-1 (Lin, 2004), the stochastic nature of metrics (e.g., G-Eval (Liu et al., 2023)) may produce undesirable fluctuations when evaluating the same model outputs.

We propose to quantify metric stability via the test-retest reliability coefficient: on the output generated by N models, we compute the metric score with the same output twice for each model. Across different models, the Pearson correlation between the two sets of scores is the test-retest reliability coefficient. One can show that this correlation is an estimate of the reliability coefficient ρ^2_{XT} as defined in Equ. 1. This is because, for the two metrics scores for a model, $X_1 = T_1 + E_1$ and $X_2 = T_1 + E_2$, the correlation in the observed scores, $\rho_{X_1X_2}$, is algebraically equivalent to $\rho^2_{X_1T_1}$, under the assumption that each model's true score doesn't change (i.e., $T_1 = T_2$) and that the amount of fluctuation in metric evaluation remains the same (i.e., $\sigma_{E_1} = \sigma_{E_2}$) across the two evaluations (see derivations in Allen and Yen, 2001).

3.1.2 Metric Consistency

Metric Consistency describes how the metric score fluctuates within a benchmark dataset, i.e., across data points. If the metric score computed on each individual data point (e.g., summarization of a specific news article) deviates more from the average score across the benchmark dataset, the metric score would be less reliable, in that it is more sensitive to perturbations in the specific data points employed in the benchmark dataset. Drawn from the estimation of internal consistency reliability in measurement theory (where a test is made up of many individual questions), the estimation of metric consistency depends on the degree to which scores from different subsets of the benchmark dataset agree with one another.

The coefficient α (Cronbach, 1951) provides a measure of metric consistency. Let J denote the total number of data points in the test dataset, Y_j the observed score (of each model) on the jth data point alone, and $X = \sum_{j=1}^{J} Y_j$ the overall score of the model on the full test set. Then α provides a lower bound to the true reliability of X, i.e.,

$$\rho^2_{XT} \geq \alpha = \frac{J}{J-1}\left[\frac{\sigma^2_X - \sum_{j=1}^{J}\sigma^2_{Y_j}}{\sigma^2_X}\right], \quad (2)$$

where $\sigma^2_{Y_j}$ is the variance of Y_j across models. Equality holds when all the individual data point scores (Y_js) have equal correlations with the true score (T), which may be violated in practice, leading to the underestimation of true reliability via the coefficient α formula.

3.2 Validity

Validity is another core component of our framework. Metrics with low validity lead to systematic measurement errors that deviate the observed score from the true score that the test purports to measure. In other word, benchmarking is valid only when the metric scores can inform their intended interpretations (e.g., model capability) and uses (e.g., predicting models' real-world behavior).

Our framework is theoretically grounded in Messick's unified theory of test validity(e.g., Messick, 1995), under which the emphasis is given to the validation of *inferences drawn* from the test score, rather than the validation of the test itself. Different types of validities should be recognized as possible ways to gather supporting evidence for the ability to draw intended inferences (interpretations and uses) from the test score. Our framework conceptualizes two types of a metric's validity, concurrent validity, and construct validity (e.g., Allen and Yen, 2001), which can be applied in different situations—when a validated reference criterion is available or not—as we elaborate below.

3.2.1 Metric Concurrent Validity

Metric Concurrent Validity relies on another validated metric as the reference criterion. This type of validity is most relevant when evaluating a metric as an alternative to existing ones that may be expensive and infeasible to acquire in practice. For example, evaluating a large number of model-generated outputs with trained human experts is often challenging, motivating the development of automatic alternatives. One can conclude that an automatic metric is a valid proxy if it has high concurrent validity using the expert valuation results as the reference criterion.

When both the target evaluation metric (X, e.g., a new automatic metric) and the reference criterion (Y, e.g., expert evaluation) are continuous, a straightforward way to quantify concurrent validity is via their Pearson correlation, ρ_{XY}, often referred to as the (criterion-related) validity coefficient. One should note that measurement error in either X or Y is expected to attenuate this correlation (Spearman, 1910): At the population level, ρ_{XY} is bounded above by the square root of the product of the two scores' (X and Y) *reliabilities*. This again highlights the importance of safeguarding the reliability of the evaluation metric, as a noisy metric with low reliability is expected to yield poor predictive power on the criterion of interest.

3.2.2 Metric Construct Validity

Construct validity, a term coined by Cronbach and Meehl, refers to the degree to which the observed behaviors on the test (e.g., test scores) can reasonably reflect the intended construct (e.g., language proficiency). This notion is directly applicable to evaluation metrics that are *explicitly* constructed to assess specific aspects of a model's performance or output quality, e.g., human evaluation (or automatic metrics, if developed specifically) on summarization, coherence, fluency, consistency, etc. However, even for metrics of which the intended construct is not explicitly defined, it is still necessary to understand what underlying dimensions of model capabilities they actually capture.

It is important to note that the underlying construct is often latent and not directly observable to assess its relation with the measure. Measurement theory, therefore, provides statistical tools to assess the construct validity of a measure through its relation with other observable variables (e.g., other tests purported to reflect the same or different constructs). We consider three such aspects of validity based on the measurement literature:

- *Metric Convergent Validity*: Whether measures of identical or related construct(s) are indeed identical or related. For example, for the same aspect of summarization quality (e.g., coherence), scores provided by different evaluation methods (e.g. by different raters) should be highly correlated.

- *Metric Divergent Validity*: Whether measures of unrelated constructs are indeed unrelated. For example, for distinct aspects of summarization quality (e.g., coherence and relevance), scores provided by the same method (e.g., by the same rater) should show substantially lower correlations than those for the same trait across methods. Low divergent validity would indicate method bias: e.g., the observed score depends greatly on the rater's subjective tendency rather than the model's performance on the rated dimension.

- *Metric Factorial Validity*: Whether the observed metric scores can be explained by a smaller number of unobserved factors. For example, if scores on multiple evaluation metrics exhibit high correlations, this might suggest the presence of a common underlying factor causing these scores to move in unison.

We introduce two statistical tools to evaluate these aspects of construct validity. Specifically, Metric Convergent Validity and Metric Divergent Validity can be evaluated through the analysis of a multitrait-multimethod (MTMM) table, and Merict Factorial Validity can be evaluated via factor analysis. Note that these validity evaluation methods will only inform if there is an underlying construct or how many of them are being captured. Defining *what* these constructs are will require further conceptualization and theorizing.

The MTMM table presents a way to scrutinize the extent to which observed test scores are indeed measuring some underlying constructs, when two or more constructs are measured using two or more methods (Campbell and Fiske, 1959). For example, when evaluating a summarization model, researchers may ask several raters to rate the generated outputs on four "traits" (aspects of output quality), e.g., coherence, consistency, fluency, and relevance. In this case, the MTMM table allows

examining whether, across different raters (evaluation methods), the raters' scores indeed appear to characterize the model's performance on four distinct constructs. By convention, an MTMM table reports the pairwise correlations of the observed metric scores across raters and traits on the off-diagonals and the reliability coefficients of each score on the diagonals. The analysis of an MTMM examines is exemplified in Sec. 4.1.2.

Factor Analysis establishes *Metric Factorial Validity* (e.g., Thurstone, 1947) by examining whether the observed metric scores can be explained by a smaller number of unobserved factors. For example, if scores from multiple evaluation metrics exhibit high correlations, this might suggest the presence of a common underlying factor causing these scores to move in unison. Under a factor analysis model, the distribution of the observed score on an indicator X_j, such as a particular evaluation metric, is a function of a linear combination of the model's factor scores on $K \geq 1$ general latent factors $(f_1, \ldots, f_K)$ and the unique score U_j on the indicator j unexplained by the latent factor, including measurement error, i.e.,

$$X_j = f(\lambda_{j1} f_1 + \ldots + \lambda_{jK} f_K + U_j). \quad (3)$$

$f(\cdot)$ can be the identity function for normally distributed observed scores, but when scores are ordinal (e.g., expert ratings on a 5-point scale) or skewed, we suggest adopting an ordinal factor model (Muthén, 1984) where $f(\cdot)$ is a step function that evaluates whether Equ. 3 exceeds specific thresholds for each score category on the latent continuum. Factor analysis can be exploratory or confirmatory. In the latter, select loadings (λ_{jk}s) are constrained to 0 to represent the theorized nomological network, e.g., an expert rating on consistency loads on no other dimensions. By establishing the *Metric Factorial Validity* through factor analysis, we could further develop more effective metrics by answering the following questions:

- Fit indices: For confirmatory factor analysis, how well does the theorized factorial structure align with the observed data?
- Factor scores: What is an NLG model's factor score on a particular dimension?
- Factor loadings: How strongly does a specific factor affect an observed evaluation metric score?
- Residual correlation: For different evaluation metrics, are the residuals (unexplained score variation by the common factors) correlated, which may suggest additional dimensions?

4 Case Study

To illustrate how to apply our framework to evaluate NLG evaluation metrics, in this section, we ran a case study on evaluating summarization metrics. As discussed, our evaluation focuses on the metrics and the results should be interpreted as dependent on the benchmark used. We leave it for future research to explore the generalizability of the results across different benchmarks.

4.1 Summarization Metric Evaluation

We conducted our analysis using the SummEval dataset (Fabbri et al., 2021), which contains a total of 1700 summaries generated by 17 models on CNN/Daily Mail benchmark dataset. In this dataset, each generated summary is rated by three experts, who provided 5-point-scale ratings on four dimensions: Coherence, Consistency, Fluency, and Relevance. We ran 16 types of popular automatic metrics that include rule-based metrics, embedding-based metrics, end-to-end metrics, and LLM-based metrics that are reference-based or reference-free, see Appx. A.1. Since the score distributions on many evaluation metrics were skewed, we normalized automatic evaluation scores for the subsequent analyses, see Appx. A.2.

4.1.1 Metric Stability and Consistency

To evaluate an automatic metric's stability, we computed the metric score twice for each model's output on each data point, calculated two sets of average scores for each model, and reported the correlation between the two sets of scores on the 17 models. A metric's consistency was evaluated via the coefficient α in Equ. 2. Fig. 2 presents the Metric Stability and Metric Consistency estimates of a set of selected metrics (full results, see Fig. 4 in Appx). We found most metrics achieved high stability. Metrics with non-deterministic algorithms, such as LLM-based metrics G-Eval, display higher levels of measurement error in terms of Metric Stability. While compared to G-Eval with GPT3.5, G-Eval with GPT-4 yields higher stability. For Metric Consistency, we found, for the ROUGE family, a longer n-gram makes the metric less reliable and more prone to potential data perturbations in the test dataset. Therefore, to mitigate measurement

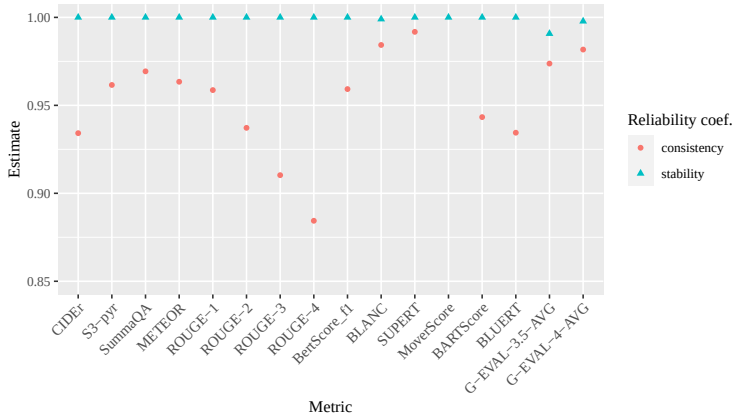

Figure 2: Estimated Metric Stability and Metric Consistency of popular NLG Metrics.

error, for less stable LLM-based metrics, the evaluator should consider applying the metrics multiple times and taking the average score and for less consistent metrics such as ROUGE-4, the evaluator should consider using a larger test dataset.

Conventionally, a reliability coefficient above .9 indicates good reliability. The metric stability and consistency estimates can help approximate the standard error of measurement of an average test dataset metric score (X), by observing from Equ. 1 that $\sigma_E = \sigma_X\sqrt{1-\rho^2_{XT}}$. For example, the sample standard deviation in the average test dataset METEOR score is .38 and the metric consistency estimate was .966, translating to an expected measurement error due to score variability across the 100 data points of $.38 \times \sqrt{1-.966} \approx .07$. A METEOR score difference between two models less than .07 would thus be of limited interest, as the difference is less than the expected amount of fluctuation in the score due to measurement error.

4.1.2 Metric Construct Validity

We begin by evaluating the construct validity of expert-based evaluation metrics included in the SumEval dataset. These evaluations were conducted in a confirmatory manner, assuming that the four ratings provided by each expert on a summarization output's Coherence, Consistency, Fluency, and Relevance indeed measure the four distinct dimensions. Table 1 presents the MTMM table for the three expert's ratings on four dimensions. Metric Convergent Validity can be examined by inspecting the bolded entries: Inter-rater agreements on the same dimension were high ($.48 - .89$) in general but lower for Relevance ($.48 - .58$). Italic entries can inform the evaluation of metric divergent validity: Overall, an expert's ratings on different dimensions showed lower correlations than for ratings by different experts on the same dimension, with the exception of Coherence and Relevance, which sometimes showed higher correlations (underscored, $.47 - .62$) than that on ratings for Relevance across raters. This may suggest that, although the expert raters were asked to separately rate on Coherence and Relevance, they might inherently be rating the summarization outputs on the same underlying characteristic.

Confirmatory factor analysis was further conducted (see Appx. A.4) to test the observed conflated validity structure indicated by the MTMM analysis. The results show that the four-factor model fitted the observed data adequately well (Comparative Fit Index $= .999$, Tucker-Lewis Index $= .999$, Root Mean Square Error of Approximation $= .047 < .05$), supporting the theorized loading structure, i.e., experts indeed rated on four factors. However, the estimated factor correlations below suggested high correlations between dimensions, especially for Coherence and Relevance. This result supports a conflated validity structure.

Since Coherence and Relevance are distinct constructs by definition (Fabbri et al., 2021), the conflated validity structure indicates potential issues in the expert rating process. One source of such systematic measurement error may come from unclear instructions or the expert rater's individual leniency. Such issues may cascade if new automatic evaluation metrics were trained or validated on this dataset.

4.2 Metric Concurrent Validity

Finally, for each automatic evaluation metric, we evaluated its concurrent validity: i.e. should the metric score be used to predict expert rating on

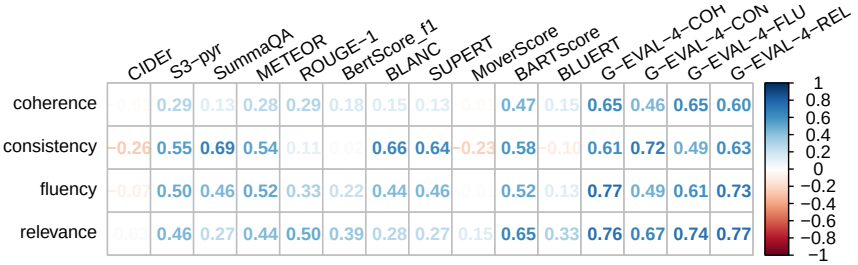

Figure 3: Concurrent validity coefficients of the selected metrics in predicting the four expert-rated dimensions' factor scores. Values are based on Kendall's τ.

		Expert 1				Expert 2				Expert 3			
		Coherence	Consistency	Fluency	Relevance	Coherence	Consistency	Fluency	Relevance	Coherence	Consistency	Fluency	Relevance
Expert 1	Coherence	0.96	*0.30*	*0.37*	*0.62*	**0.71**	0.30	0.30	0.51	**0.60**	0.28	0.29	0.38
	Consistency	-	0.98	*0.44*	*0.30*	0.27	**0.89**	0.40	0.40	0.22	**0.89**	0.42	0.23
	Fluency	-	-	0.97	*0.31*	0.37	0.45	**0.72**	0.36	0.31	0.41	**0.74**	0.24
	Relevance	-	-	-	0.93	0.49	0.29	0.23	**0.58**	0.41	0.29	0.26	**0.48**
Expert 2	Coherence	-	-	-	-	0.99	*0.30*	*0.31*	*0.59*	**0.62**	0.27	0.28	0.37
	Consistency	-	-	-	-	-	0.98	*0.41*	*0.43*	0.23	**0.94**	0.44	0.25
	Fluency	-	-	-	-	-	-	0.97	*0.29*	0.23	0.39	**0.72**	0.19
	Relevance	-	-	-	-	-	-	-	0.98	0.43	0.42	0.32	**0.50**
Expert 3	Coherence	-	-	-	-	-	-	-	-	0.96	*0.21*	*0.24*	*0.47*
	Consistency	-	-	-	-	-	-	-	-	-	0.98	*0.44*	*0.24*
	Fluency	-	-	-	-	-	-	-	-	-	-	0.96	*0.20*
	Relevance	-	-	-	-	-	-	-	-	-	-	-	0.90

Table 1: Multitrait-Multimethod table of the pairwise correlations between expert ratings.
Notes: Entries in bold are the correlations of ratings on the same dimension by different experts. Entries in italic are the correlations of the ratings on different dimensions by the same expert. Underscored entries are the correlations between coherence and relevance ratings by the same expert, which showed strong correlations.

each dimension as a more cost-efficient alternative? Different from prior studies (Fabbri et al., 2021), we report Kendall's rank correlations between each model's metric scores and the factor scores (instead of the raw means) based on expert ratings on the four dimensions. The metric concurrent validity coefficients are presented in Fig. 3 (for full results, see Fig.7 in Appx.).

In this analysis, we found that although BARTScore and G-Eval are sensitive to detecting quality signals in all four expert-rated dimensions, the lack of variance in the validity coefficients surfaces another issue—their lack of capability to discriminative different dimensions. For example, although G-EVAL-4-COH is designed for Coherence, it strongly correlates with Fluency and Relevance, which means it is unable to detect Coherence differences if two models are both strong at Fluency and Relevance. On the contrary, SummaQA only reacts to Consistency which makes it a more desirable metric even though its correlation with expert rating is lower than G-EVAL-4-CON in general, if the evaluation objective is Consistency.

5 Related Work

To evaluate NLG evaluation metrics, NLG communities have employed various methods. One of the most commonly used methods is by correlation with human judgment. However, the inconsistency and subjectivity of human judgment when rating generated content, in addition to the non-transparent and non-standardized annotation process (Sai et al., 2021; Liu et al., 2016; Reiter, 2018; Celikyilmaz et al., 2020; Kauchak and Barzilay, 2006), create a shaking foundation when it is used to validate NLG metrics. Other methods have also been applied to evaluate evaluation metrics, including example-based qualitative analysis (Tao et al., 2018), perturbation (Sai et al., 2021), and meta-analysis (Reiter, 2018). Although qualitative analyses provide in-depth insights, the challenge of scalability limits their capability to capture a comprehensive picture of metric quality.

6 Conclusion

Evaluation metrics guide model development. Drawing from the core concept of *reliability* and *validity* in measurement theory, we present a Metric Evaluation Framework that defines and operationalizes four key desiderata for NLG metrics. With a collection of statistical tools, our framework offers the NLP community an effective and principled way to analyze, evaluate and understand NLG evaluation metrics.

7 Limitation

We recognize the following limitations of our paper. First, evaluating evaluation metrics for NLG models should not be treated as a single-shot task. Instead, as suggested in Messick's unified theory of validity (Messick, 1995), it is essential to continuously gather cumulative evidence of validity to ensure the ongoing effectiveness and reliability of the metrics. The process of accumulating valid evidence is an iterative and dynamic endeavor that aligns with the evolving landscape of NLG models and their applications. Future studies are necessary to collect other types of evidence, such as a metric's ability to predict users' sanctification, to continuously evaluate the effectiveness of an NLG metric.

Second, measurement errors may surface and accumulate at every stage of the evaluation process, including benchmark design, data collection, etc. To perform the analysis of evaluation metrics, we have to assume the reliability and validity of the other parts of the evaluation process. Therefore, the results of the case study should be interpreted as dependent on the benchmark used, e.g., CNN/Daily Mail dataset. Future study is required to study the generalizability of the results across different benchmarks.

Third, our framework does not aim to provide comprehensive coverage of all of measurement error sources in NLG evaluation metrics. For example, we did not discuss predictive validity in our framework despite its importance in education and psychological testing. We encourage researchers and practitioners to extend our framework for other types of reliability and validity and build datasets to support more comprehensive analysis, e.g., a dataset with the model's real-world performance, to deepen our knowledge of NLG metric evaluation.

References

Mary J Allen and Wendy M Yen. 2001. *Introduction to measurement theory*. Waveland Press.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Anja Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021. The reprogen shared task on reproducibility of human evaluations in nlg: Overview and results. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 249–258.

Donald T Campbell and Donald W Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, 56(2):81.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's 'human'is not gold: Evaluating human evaluation of generated text. In *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)*, pages 7282–7296.

Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334.

Lee J Cronbach and Paul E Meehl. 1955. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. Supert: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. *arXiv preprint arXiv:2005.03724*.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *arXiv preprint arXiv:2202.06935*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

David Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definition. Association for Computational Linguistics (ACL).

David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 455–462.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A Smith, and Daniel S Weld. 2021. Genie: Toward reproducible and standardized human evaluation for text generation. *arXiv preprint arXiv:2101.06561*.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9):741.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Bengt Muthén. 1984. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1):115–132.

Preksha Nema and Mitesh M Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3950–3959.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. *arXiv preprint arXiv:1508.06034*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Ananya B Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M Khapra. 2021. Perturbation checklists for evaluating nlg evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. *arXiv preprint arXiv:1909.01610*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. *arXiv preprint arXiv:2105.04054*.

C Spearman. 1904. General intelligence, objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292.

Charles Spearman. 1910. Correlation calculated from faulty data. *British journal of psychology*, 3(3):271.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Louis Leon Thurstone. 1947. Multiple-factor analysis; a development and expansion of the vectors of mind.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the blanc: Human-free quality estimation of document summaries. *arXiv preprint arXiv:2002.09836*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Catherine Yeo and Alyssa Chen. 2020. Defining and evaluating fair natural language generation. *arXiv preprint arXiv:2008.01548*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. Deconstructing nlg evaluation: Evaluation practices, assumptions, and their implications. *arXiv preprint arXiv:2205.06828*.

A Appendix

A.1 Metrics

Our selection of evaluation methods includes popular metrics for NLG tasks including both reference-based and reference-free metrics. Compared to the original SummEval dataset, we additionally selected end-to-end metrics and recent LLM-based metrics.

ROUGE (Lin, 2004) evaluates the generated summary by comparing the number of overlapping word sequences (n-grams) with a set of reference summaries.

ROUGE-WE (Ng and Abrecht, 2015) expands on ROUGE by incorporating soft lexical matching, which utilizes the cosine similarity of Word2Vec (Mikolov et al., 2013) embeddings.

S3 (Peyrard et al., 2017) is a model-based metric that combines existing evaluation metrics like ROUGE, JS-divergence, and ROUGE-WE. It utilizes these metrics as input features to predict the evaluation score.

BertScore (Zhang et al., 2019) calculates similarity scores by aligning the generated and reference summaries at the token-level. Token alignments are determined greedily to maximize the cosine similarity between contextualized token embeddings from BERT (Devlin et al., 2018).

MoverScore (Zhao et al., 2019) quantifies the semantic distance between a summary and a reference text by utilizing the Word Mover's Distance (Kusner et al., 2015). This distance measure operates over n-gram embeddings obtained from BERT representations.

SummaQA (Scialom et al., 2019) utilizes a BERT-based question-answering model to respond to cloze-style questions using generated summaries. This metric provides both the F1 overlap score and the confidence of the QA model.

BLANC (Vasilyev et al., 2020) is a reference-less metric that assesses the performance improvement of a pre-trained language model when provided with a document summary while performing language understanding tasks on the original document's text.

SUPERT (Gao et al., 2020) is a reference-less metric that measures the semantic similarity between model outputs and pseudo-reference summaries generated by extracting significant sentences from the source documents using soft token alignment techniques.

BLEU (Papineni et al., 2002) is a metric that focuses on precision at the corpus level. It calculates the n-gram overlap between a candidate utterance and a reference utterance while incorporating a penalty for brevity.

CHRF (Popović, 2017) measures character-based n-gram overlap between model outputs and reference documents.

METEOR (Banerjee and Lavie, 2005) determines an alignment between candidate and reference sentences by mapping unigrams in the generated summary to 0 or 1 unigrams in the reference, taking into account stemming, synonyms, and paraphrases.

CIDer (Vedantam et al., 2015) calculates the co-occurrence of 1-4 gram units between the candidate and reference texts, giving less weight to common n-grams and computing cosine similarity between the n-grams of the candidate and reference texts.

BARTScore (Yuan et al., 2021) evaluates text directly based on the probability of being generated from or generating other outputs. It addresses the modeling challenge using a pre-trained sequence-to-sequence (seq2seq) model called BART (Lewis et al., 2019).

BLEURT (Sellam et al., 2020) is a BERT-based metric that can model human judgments with a few thousand training examples, which may introduce some bias.

G-Eval (Liu et al., 2023) is a framework that leverages LLM with Chain-of-Thoughts (CoT) (Wei et al., 2022) to evaluate the quality of generated text. The generated outputs are assessed using a set of prompts along with generated CoT.

Data Statistics (Grusky et al., 2018) define three measures of dataset extractiveness: extractive fragment coverage, density, and compression ratio. Extractive fragment coverage quantifies the percentage of words in the summary that are derived from the source article, indicating the degree to which the summary is a derivative of the original text. Density represents the average length of the extractive fragment to which each summary word belongs. Compression ratio measures the word ratio between the articles and their summaries.

A.2 Metric Normalization

Initial exploratory analysis revealed that the score distributions on many evaluation metrics were skewed. We thus normalized each automatic evaluation score (via the transformation $X_j^* = \Phi^{-1}(\frac{1}{N}\sum_{i=1}^{N}\mathcal{I}(X_{ij} \leq X_j))$) and subsequently worked with normalized automatic metric scores, which approximately followed $N(0,1)$ distribution and are more appropriate for correlational analysis and linear models.

A.3 Metric Stability and Consistency Results

Fig. 4 presents the Metric Stability and Metric Consistency estimates of all automatic evaluations and the metric consistency estimates of all expert and automatic metrics.

A.4 Confirmatory Factor Analysis on Expert Ratings

Confirmatory factor analysis was further conducted on the 12 (3x4) expert ratings, assuming that each rating loads only on the corresponding dimension. Given that the expert ratings were highly skewed (see Fig. 5 in Appx.), an ordinal factor model (Muthén, 1984) was fitted. Judging from commonly used fit indices (Comparative Fit Index $= .999$, Tucker-Lewis Index $= .999$, Root Mean Square Error of Approximation $= .047 < .05$), the four-factor model fitted the observed data adequately well, supporting the theorized loading structure, e.g., Experts rated on four factors. Tab 2 in Appx reports the estimated factor loadings and thresholds of each expert rating, assuming that the four latent factors each have mean 0 and SD of 1. Loadings were generally high (adopting the $\geq .4$ convention) but varied across experts and dimensions (generally lower for Relevance). Rater differences were also found in their leniency: For instance, expert 3 was more likely than experts 1 and 2 to provide a rating of 5 (with lower threshold estimates for score 5) on output Consistency, but less likely to do so (with higher threshold estimates) on Coherence and Relevance. The estimated factor correlations below suggested high correlations between dimensions, especially for coherence and relevance:

	Coherence	Consistency	Fluency
Consistency	.51	–	–
Fluency	.56	.68	–
Relevance	.86	.64	.53

A.5 Residual Analysis

We further performed principal component analysis on the residuals of the automatic evaluations, which capture the unexplained variance by the 4 dimensions' factor scores. Plot of the first two principal components is shown in Fig. 6. Here, visual clusters of evaluation metrics are found, suggesting that select metrics likely tapped on common additional dimensions. The unexplained residual variance may guide future investigation on discovering other quality signals in summarization tasks.

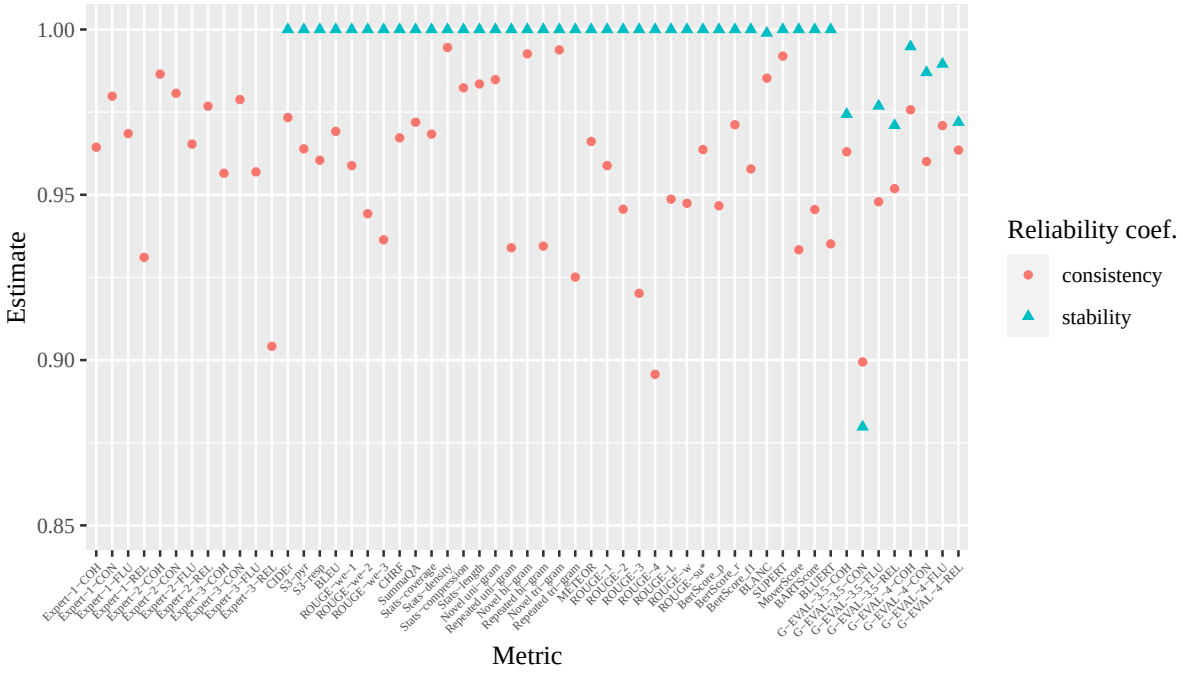

Figure 4: Metric stability and consistency estimates for all expert- and metric-based scores.

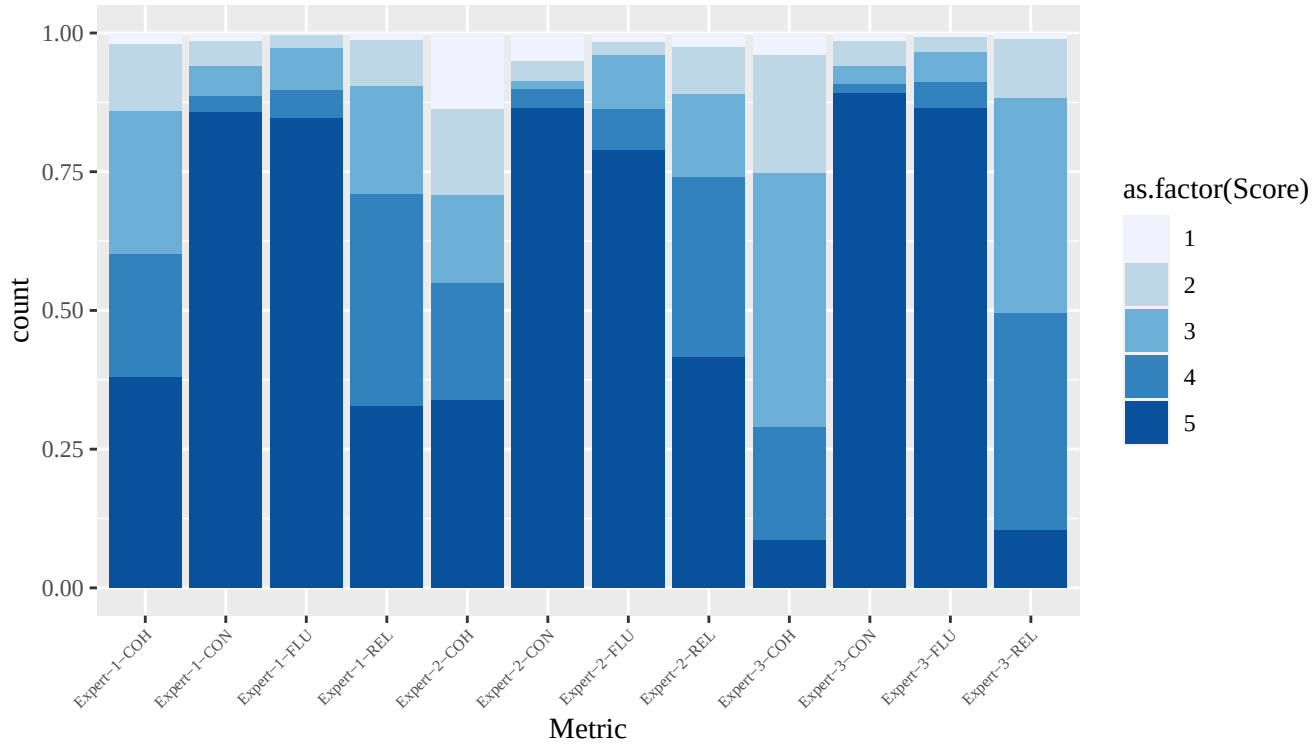

Figure 5: Distribution of 5-point scale expert ratings.

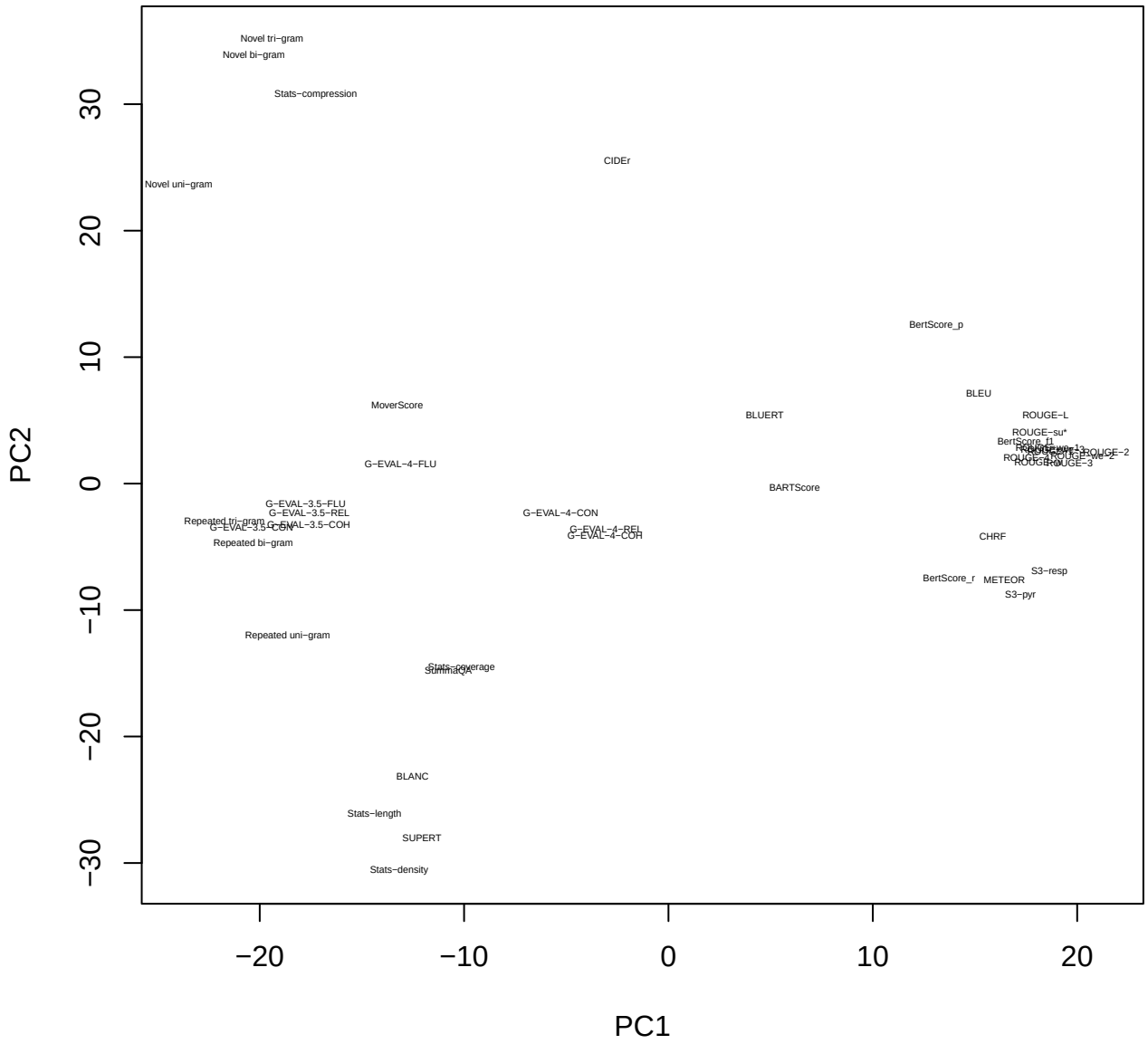

Figure 6: Plot of the first two principal components of the residuals of the 4-factor model.

Table 2: Factor loading and score category threshold estimates for the 4-factor confirmatory model of ordinal expert ratings.

	Loading				Threshold			
	coherence	consistency	fluency	relevance	2	3	4	5
expert_1_coherence	0.90	-	-	-	-2.03	-1.08	-0.26	0.31
expert_2_coherence	0.88	-	-	-	-1.09	-0.55	-0.12	0.42
expert_3_coherence	0.76	-	-	-	-1.74	-0.67	0.56	1.37
expert_1_consistency	-	0.97	-	-	-2.15	-1.56	-1.20	-1.07
expert_2_consistency	-	0.98	-	-	-1.63	-1.36	-1.27	-1.10
expert_3_consistency	-	1.00	-	-	-2.16	-1.55	-1.33	-1.23
expert_1_fluency	-	-	0.98	-	-2.60	-1.93	-1.26	-1.02
expert_2_fluency	-	-	0.88	-	-2.12	-1.74	-1.09	-0.80
expert_3_fluency	-	-	0.92	-	-2.49	-1.81	-1.35	-1.10
expert_1_relevance	-	-	-	0.79	-2.25	-1.30	-0.55	0.45
expert_2_relevance	-	-	-	0.85	-1.95	-1.23	-0.65	0.22
expert_3_relevance	-	-	-	0.64	-2.28	-1.19	0.01	1.26

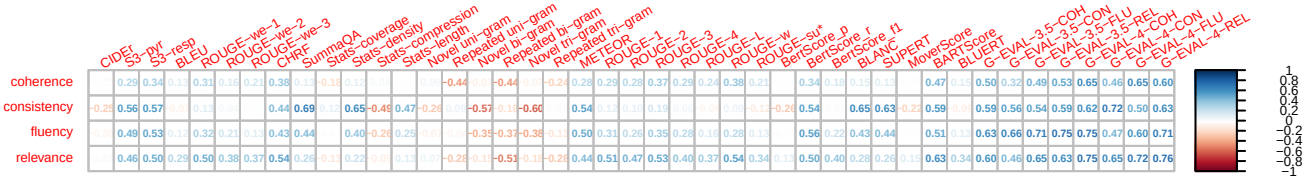

Figure 7: Concurrent validity coefficients of the metric-based scores in predicting the four expert-rated dimensions' factor scores. Values are based on Kendall's τ.