Quantified Reproducibility Assessment of NLP Results

Anya Belz and **Maja Popović**
ADAPT Research Centre
Dublin City University, Ireland
{anya.belz,maja.popovic}@adaptcentre.ie

Simon Mille
Universitat Pompeu Fabra
Barcelona, Spain
simon.mille@upf.edu

Abstract

This paper describes and tests a method for carrying out quantified reproducibility assessment (QRA) that is based on concepts and definitions from metrology. QRA produces a single score estimating the degree of reproducibility of a given system and evaluation measure, on the basis of the scores from, and differences between, different reproductions. We test QRA on 18 system and evaluation measure combinations (involving diverse NLP tasks and types of evaluation), for each of which we have the original results and one to seven reproduction results. The proposed QRA method produces degree-of-reproducibility scores that are comparable across multiple reproductions not only of the same, but of different original studies. We find that the proposed method facilitates insights into causes of variation between reproductions, and allows conclusions to be drawn about what changes to system and/or evaluation design might lead to improved reproducibility.

1 Introduction

Reproduction studies are becoming more common in Natural Language Processing (NLP), with the first shared tasks being organised, including REPROLANG (Branco et al., 2020) and ReproGen (Belz et al., 2021b). In NLP, reproduction studies generally address the following question: *if we create and/or evaluate this system multiple times, will we obtain the same results?*

To answer this question for a given specific system, typically (Wieling et al., 2018; Arhiliuc et al., 2020; Popović and Belz, 2021) an original study is selected and repeated more or less closely, before comparing the results obtained in the original study with those obtained in the repeat, and deciding whether the two sets of results are similar enough to support the same conclusions.

This framing, whether the same conclusions can be drawn, involves subjective judgments and different researchers can come to contradictory conclusions: e.g. the four papers (Arhiliuc et al., 2020; Bestgen, 2020; Caines and Buttery, 2020; Huber and Çöltekin, 2020) reproducing Vajjala and Rama (2018) in REPROLANG all report similarly large differences, but only Arhiliuc et al. conclude that reproduction was unsuccessful.

There is no standard way of going about a reproduction study in NLP, and different reproduction studies of the same original set of results can differ substantially in terms of their similarity in system and/or evaluation design (as is the case with the Vajjala and Rama (2018) reproductions, see Section 4 for details). Other things being equal, a more similar reproduction can be expected to produce more similar results, and such (dis)similarities should be factored into reproduction analysis and conclusions, but NLP lacks a method for doing so.

Being able to assess reproducibility of results objectively and comparably is important not only to establish that results are valid, but to provide evidence about which methods have better/worse reproducibility and what may need to be changed to improve reproducibility. To do this, assessment has to be done in a way that is also comparable across reproduction studies of different original studies, e.g. to develop common expectations of how similar original and reproduction results should be for different types of system, task and evaluation.

In this paper, we (i) describe a method for quantified reproducibility assessment (QRA) directly derived from standard concepts and definitions from metrology which addresses the above issues, and (ii) test it on diverse sets of NLP results. Following a review of related research (Section 2), we present the method (Section 3), tests and results (Section 4), discuss method and results (Section 5), and finish with some conclusions (Section 6).

2 Related Research

The situation memorably caricatured by Pedersen (2008) still happens all the time: you download

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 16 - 28
May 22-27, 2022 ©2022 Association for Computational Linguistics

some code you read about in a paper and liked the sound of, you run it on the data provided, only to find that the results are not the same as reported in the paper, in fact they are likely to be worse (Belz et al., 2021a). When both data and code are provided, the number of potential causes of such differences is limited, and the NLP field has shared increasingly detailed information about system, dependencies and evaluation to chase down sources of differences. Sharing code and data together with detailed information about them is now expected as standard, and checklists and datasheets have been proposed to standardise information sharing (Pineau, 2020; Shimorina and Belz, 2021).

Reproducibility more generally is becoming more of a research focus. There have been several workshops and initiatives on reproducibility, including workshops at ICML 2017 and 2018, the reproducibility challenge at ICLR 2018 and 2019, and at NeurIPS 2019 and 2020, the REPROLANG (Branco et al., 2020) initiative at LREC 2020, and the ReproGen shared task on reproducibility in NLG (Belz et al., 2021b).

Despite this growing body of research, no consensus has emerged about standards, terminology and definitions. Particularly for the two most frequently used terms, *reproducibility* and *replicability*, multiple divergent definitions are in use, variously conditioned on same vs. different teams, methods, artifacts, code, and data. For example, for Rougier et al. (2017), *reproducing* a result means running the same code on the same data and obtaining the same result, while *replicating* the result is writing and running new code based on the information provided by the original publication. For Wieling et al. (2018), *reproducibility* is achieving the same results using the same data and methods.

According to the ACM's definitions (Association for Computing Machinery, 2020), results have been *reproduced* if obtained in a different study by a different team using artifacts supplied in part by the original authors, and *replicated* if obtained in a different study by a different team using artifacts *not* supplied by the original authors. The ACM originally had these definitions the other way around until asked by ISO to bring them in line with the scientific standard (ibid.).

Conversely, in Drummond's view 2009 obtaining the same result by re-running an experiment in the same way as the original is *replicability*, while *reproducibility* is obtaining it in a different way. Whitaker (2017), followed by Schloss (2018), defines four concepts rather than two, basing definitions of *reproducibility, replicability, robustness* and *generalisability* on the different possible combinations of same vs. different data and code.

None of these definitions adopt the general scientific concepts and definitions pertaining to reproducibility, codified in the International Vocabulary of Metrology, VIM (JCGM, 2012). One issue is that they all reduce the in principle open-ended number of dimensions of variation between measurements accounted for by VIM to just two or three (code, data and/or team). Another, that unlike VIM, they don't produce comparable results.

NLP does not currently have a shared approach to deciding reproducibility, and results from reproductions as currently reported are not comparable across studies and can, as mentioned in the introduction, lead to contradictory conclusions about an original study's reproducibility. There appears to be no work at all in NLP that aims to estimate *degree* of reproducibility which would allow cross-study comparisons and conclusions.

3 Metrology-based Reproducibility Assessment

Metrology is a meta-science: its subject is the standardisation of measurements across all of science to ensure comparability. Computer science has long borrowed terms, most notably reproducibility, from metrology, albeit not adopting the same definitions (as discussed in Section 2 above).

In this section, we describe quantified reproducibility assessment (QRA), an approach that is directly derived from the concepts and definitions of metrology, adopting the latter exactly as they are, and yields assessments of the degree of similarity between numerical results and between the studies that produced them. We start below with the concepts and definitions that QRA is based on, followed by an overview of the framework (Section 3.2) and steps in applying it in practice (Section 3.3).

3.1 VIM Definitions of Repeatability and Reproducibility

The International Vocabulary of Metrology (VIM) (JCGM, 2012) defines repeatability and reproducibility as follows (defined terms in bold, see VIM for subsidiary defined terms):

2.21 **measurement repeatability** (or repeatability, for short) is **measurement precision** under a set of **repeatability conditions of measurement**.

2.20 a **repeatability condition of measurement** (repeatability condition) is a condition of **measurement**, out of a set of conditions that includes the same **measurement procedure**, same operators, same **measuring system**, same operating conditions and same location, and replicate measurements on the same or similar objects over a short period of time.

2.25 **measurement reproducibility** (reproducibility) is **measurement precision** under **reproducibility conditions of measurement**.

2.24 a **reproducibility condition of measurement** (reproducibility condition) is a condition of **measurement**, out of a set of conditions that includes different locations, operators, **measuring systems**, etc. A specification should give the conditions changed and unchanged, to the extent practical.

In other words, VIM considers repeatability and reproducibility to be properties of measurements (not objects, scores, results or conclusions), and defines them as measurement precision, i.e. both are quantified by calculating the precision of a set of measured quantity values. Both concepts are defined relative to a set of conditions of measurement: the conditions have to be known and specified for assessment of repeatability and reproducibility to be meaningful. In repeatability, conditions are the same, whereas in reproducibility, they differ.

In an NLP context, objects are systems, and measurements involve applying an evaluation method to a system usually via obtaining a sample of its outputs and applying the method to the sample (further details of how concepts map to NLP are provided in Section 3.3).

3.2 Assessment framework

The VIM definitions translate directly to the following definition of repeatability R^0 (where all conditions of measurement C are the same across measurements):

$$\begin{aligned} R^0(M_1, M_2, ...M_n) := \text{Precision}(v_1, v_2, ...v_n), \\ \text{where } M_i\colon (m, O, t_i, C) \mapsto v_i \end{aligned} \tag{1}$$

and the M_i are repeat measurements for measurand m performed on object O at different times t_i under (the same) set of conditions C, producing measured quantity values v_i. Below, the coefficient of variation is used as the precision measure, but other measures are possible. Conditions of measurement are attribute/value pairs each consisting of a name and a value (for examples, see following section). Reproducibility R is defined in the same way as R^0 except that condition *values* (but not names) differ for one or more of the conditions of measurement C_i:

$$\begin{aligned} R(M_1, M_2, ...M_n) := \text{Precision}(v_1, v_2, ...v_n), \\ \text{where } M_i\colon (m, O, t_i, C_i) \mapsto v_i \end{aligned} \tag{2}$$

Precision is typically reported in terms of some or all of the following: mean, standard deviation with 95% confidence intervals, coefficient of variation, and percentage of measured quantity values within n standard deviations. We opt for the coefficient of variation (CV),[1] because it is a general measure, not in the unit of the measurements (unlike mean and standard deviation), providing a quantification of precision (degree of reproducibility) that is comparable across studies (Ahmed, 1995, p. 57). This also holds for percentage within n standard deviations but the latter is a less recognised measure, and likely to be the less intuitive for many.

In reproduction studies in NLP/ML, sample sizes tend to be very small (a sample size of 8, one original study plus 7 reproductions, as in Table 6 is currently unique). We therefore need to use de-biased sample estimators: we use the unbiased sample standard deviation, denoted s^*, with confidence intervals calculated using a t-distribution, and standard error (of the unbiased sample standard deviation) approximated on the basis of the standard error of the unbiased sample variance $\text{se}(s^2)$ as $\text{se}_{s^2}(s^*) \approx \frac{1}{2\sigma}\text{se}(s^2)$ (Rao, 1973). Assuming measured quantity values are normally distributed, we calculate the standard error of the sample variance in the usual way: $\text{se}(s^2) = \sqrt{\frac{2\sigma^4}{n-1}}$. Finally, we also use a small sample correction (indicated by the star) for the coefficient of variation: $\text{CV}^* = (1 + \frac{1}{4n})\text{CV}$ (Sokal and Rohlf, 1971).[2]

Before applying CV* to values on scales that do not start at 0 (mostly in human evaluations) we shift values to start at 0 to ensure comparability.[3] This means that to calculate the CV* scores in the tables below, measurements are first shifted.

[1] The coefficient of variation (CV), also known as relative standard deviation (RSD) is defined as the standard deviation over the mean, often expressed as a percentage.

[2] Code and data are available here: https://github.com/asbelz/coeff-var.

[3] Otherwise CV* reflects differences solely due to different lower ends of scales.

3.3 Application of the framework

Using the defined VIM terms and the notations from Section 3.2, we can refine the question from the start of this paper as follows: *if we perform multiple measurements of object* O *and measurand* m *under reproducibility conditions of measurement* C_i, *what is the precision of the measured quantity values we obtain?* For NLP, this means calculating the precision of multiple evaluation scores for the same system and evaluation measure.

Focusing here on reproducibility assessment where we start from an existing set of results (rather than a set of experiments specifically designed to test reproducibility), the steps in performing QRA are as follows:

1. For a set of n measurements to be assessed, identify the shared object and measurand.
2. Identify all conditions of measurement C_i for which information is available for all measurements, and specify values for each condition, including measurement method and procedure.
3. Gather the n measured quantity values $v_1, v_2, ...v_n$.
4. Compute precision for $v_1, v_2, ...v_n$, giving reproducibility score R.
5. Report resulting R score and associated confidence statistics, alongside the C_i.

In NLP terms, the object is the ready-to-use system (binaries if available; otherwise code, dependencies, parameter values, how the system was compiled and trained) being evaluated (e.g. the NTS-default system variant in Table 1), the measurand is the quantity intended to be measured (e.g. BLEU-style modified n-gram precision), and measurement method and procedure capture how to evaluate the system (e.g. obtaining system outputs for a specified set of inputs, and applying preprocessing and a given BLEU implementation to the latter).

VIM holds that reproducibility assessment is only meaningful if the reproducibility conditions of measurement are specified for a given test. Conditions of measurement cover every aspect and detail of how a measurement was performed and how the measured quantity value was obtained. The key objective is to capture all respects in which the measurements to be assessed are *known* to be either the same or different. If QRA is performed for a set of existing results, it is often not possible to discover every aspect and detail of how a measurement was performed, so a reduced set may have to be used (unlike in experiments designed to test reproducibility where such details can be gathered as part of the experimental design).

The reproducibility and evaluation checklists mentioned in Section 2 (Pineau, 2020; Shimorina and Belz, 2021) capture properties that are in effect conditions of measurement, and in combination with code, data and other resources serve well as a way of specifying conditions of measurement, *if* they have been completed by authors. However, at the present time, completed checklists are not normally available. The following is a simple set of conditions of measurement the information required for which *is* typically available for existing work (we include object and measurand for completeness although strictly they are not conditions, as they must be the same in each measurement in a given QRA test):

1. **Object**: the system (variant) being evaluated.[4] *E.g. a given MT system.*
2. **Measurand**: the quantity intended to be evaluated.[5] *E.g. BLEU-style n-gram precision or human-assessed Fluency.*
3. Object conditions:
   (a) **System code**: source code including any parameters. *E.g. the complete code implementing an MT system.*
   (b) **Compile/training information**: steps from code plus parameters to fully compiled and trained system, including dependencies and environment. *E.g. complete information about how the MT system code was compiled and the system trained.*
4. Measurement method conditions:[6]
   (a) **Method specification**: full description of method used for obtaining values quantifying the measurand. *E.g. a formal definition of BLEU.*
   (b) **Implementation**: the method implemented in a form that can be applied to the object in order to obtain measured quantity values. *E.g. a full implementation of BLEU.*

[4] VIM doesn't define 'object' but refers to it as that which is being measured.

[5] For definition of 'measurand' see VIM 2.3.

[6] For definition of 'measurement method', see VIM 2.5.

System (Object)	Evaluation measure (Measurand)	N scores	Papers reporting results	NLP task	Evaluation type
PASS	Clarity	2	van der Lee et al. (2017), Mille et al. (2021)	data-to-text	human, intrinsic
	Fluency	2			
	Identifiability of stance	2			
mult-base	wf1	8	Vajjala and Rama (2018), Huber and Çöltekin (2020), Arhiliuc et al. (2020), Bestgen (2020), Caines and Buttery (2020)	multilingual essay scoring as text classification	metric: intrinsic, evaluated against single reference
mult-word$^-$	wF1	8			
mult-word$^+$	wF1	8			
mult-POS$^-$	wF1	8			
mult-POS$^+$	wF1	8			
mult-dep$^-$	wF1	8			
mult-dep$^+$	wF1	8			
mult-dom$^-$	wF1	8			
mult-dom$^+$	wF1	8			
mult-emb$^-$	wF1	8			
mult-emb$^+$	wF1	8			
NTS_default	BLEU	7	Nisioi et al. (2017), Cooper & Shardlow (2020), additional reproduction study for this paper	text simplification	metric: intrinsic, eval. against input and/or multiple references
	SARI	5			
NTS-w2v_default	BLEU	6			
	SARI	4			

Table 1: Summary overview of the 18 object/measurand combinations taht were QRA-tested for this paper.

5. Measurement procedure conditions:[7]
   (a) **Procedure**: specification of how system outputs (or other system characteristics) are obtained and the measurement method is applied to them. *E.g. running a BLEU tool on system outputs and reference outputs.*
   (b) **Test set**: the data used in obtaining and evaluating system outputs (or other system characteristics). *E.g. a test set of source-language texts and reference translations.*
   (c) **Performed by**: who performed the measurement procedure and any additional information about how they did it. *E.g. the team applying the BLEU tool, and the run-time environment they used.*

The *names* of the conditions of measurement used in this paper are boldfaced above. The *values* for each condition characterise how measurements differ in respect of the condition. In reporting results from QRA tests in the following section, we use paper identifiers as shorthand for each distinct condition value (full details in each case being available from the referenced papers).

4 QRA Tests

Table 1 provides an overview of the 18 object/ measurand pairs (corresponding to 116 individual measurements) for which we performed QRA tests in this study. For each object/measurand pair, the columns show, from left to right, information about the system evaluated (object), the evaluation measure applied (measurand), the number of scores (measured quantity values) obtained, the papers in which systems and scores were first reported, and the NLP task and type of evaluation involved.

There are three sets of related systems: (i) the (single) PASS football report generator (van der Lee et al., 2017), (ii) Vajjala and Rama (2018)'s 11 multilingual essay scoring system variants, and (iii) two variants of Nisioi et al. (2017)'s neural text simplifier (NTS). PASS is evaluated with three evaluation measures (human-assessed Clarity, Fluency and Stance Identifiability), the essay scoring systems with one (weighted F1), and the NTS systems with two (BLEU and SARI). For PASS we have one reproduction study, for the essay scorers seven, and for the NTS systems, from three to six. The PASS reproduction was carried out as part of ReproGen (Belz et al., 2021b), the reproductions of the essay-scoring systems and of one of the NTS systems as part of REPROLANG (Branco et al., 2020), and we carried out an additional reproduction study of the NTS systems for this paper.[8]

The PASS text generation system is rule-based, the essay classifiers are 'theory-guided and data-driven' hybrids, and the text simplifiers are end-to-end neural systems. This gives us a good breadth

[7] For definition of 'measurement procedure', see VIM 2.6.

[8] Authors of original studies gave permission for their work to be reproduced (Branco et al., 2020; Belz et al., 2021b).

Object	Measurand	Measured quantity value		Sample size	mean	stdev	stdev 95% CI	CV* ↓
		van der Lee et al. (2017)	Mille et al. (2021)					
PASS	Clarity	5.64	6.30	2	4.969	0.583	[-2.75, 3.92]	13.193
	Fluency	5.36	6.14	2	4.75	0.691	[-3.26, 4.65]	16.372
	Stance id.	91%	97%	2	93.88	5.096	[-24.05, 34.24]	6.107

Table 2: Precision (CV*) and component measures (mean, standard deviation, standard deviation, confidence intervals) for measured quantity values obtained in two measurements for each of the three human-assessed evaluation measures for the **PASS system**. Columns 6–9 calculated on shifted scores (see Section 3.2).

Object	Measurand	*Object conditions*		*Measurement method conditions*		*Measurement procedure conditions*			Measured quantity value	*CV**
		Code by	*Comp./trained by*	*Method*	*Implem. by*	*Procedure*	*Test set*	*Performed by*		
PASS	Clarity	vdL&al	vdL&al	vdL&al	vdL&al	vdL&al	vdL&al	vdL&al	5.64	*13.193*
		vdL&al	vdL&al	vdL&al	M&al	M&al	vdL&al	M&al	6.30	
	Fluency	vdL&al	vdL&al	vdL&al	vdL&al	vdL&al	vdL&al	vdL&al	5.36	*16.372*
		vdL&al	vdL&al	vdL&al	M&al	M&al	vdL&al	M&al	6.14	
	Stance id.	vdL&al	vdL&al	vdL&al	vdL&al	vdL&al	vdL&al	vdL&al	91%	*6.107*
		vdL&al	vdL&al	vdL&al	M&al	M&al	vdL&al	M&al	96.75%	

Table 3: Conditions of measurement for two measurements each for three evaluation measures (measurands) and the **PASS system**. vdL&al = van der Lee et al. (2017); M&al = Mille et al. (2021).

of NLP tasks, system types, and evaluation types and measures to test QRA on.

4.1 QRA for NTS systems

The neural text simplification systems reported by Nisioi et al. (2017) were evaluated with BLEU (n-gram similarity between outputs and multiple reference texts) and SARI (based on word added/retained/deleted in outputs compared to both inputs and reference texts, summing over addition and retention F-scores and deletion Precisions).

Table 4 shows BLEU and SARI scores for the two system variants from the original paper and the two reproduction studies, alongside the four corresponding CV* values. In their reproduction, Cooper and Shardlow (2020) regenerated test outputs for NTS-w2v_def, but not for NTS_def, which explains the missing scores in Column 4. The different numbers of scores in different rows in Columns 6–9 are due to our own reproduction using Nisioi et al.'s SARI script, but two different BLEU scripts: (i) Nisioi et al.'s script albeit with the tokeniser replaced by our own because the former did not work due to changes in the NLTK library; and (ii) SacreBLEU (Xu et al., 2016).

Table 5 shows the conditions of measurement for each of the 22 individual measurements. The measured quantity values for those measurements where *Comp./trained by=Nisioi et al.* are identical for the SARI metric (scores highlighted by green/lighter shading and italics), but differ by up to 1.4 points for BLEU (scores highlighted by blue/darker shading). Because *Test set=Nisioi et al.* in all cases, the differences in these BLEU scores can only be caused by differences in BLEU scripts and how they were run. The corresponding CV* is as big as 0.838 for (just) the four NTS_def BLEU scores, and 1.314 for (just) the three NTS-w2v_def BLEU scores, reflecting known problems with non-standardised BLEU scripts (Post, 2018).

If we conversely look just at those measurements (identifiable by boldfaced measured quantity values in Table 5) where the reproducing team regenerated outputs (with the same system code) and evaluation scripts were the same, SARI CV* is 3.11 for the NTS_def variants, and 4.05 for the NTS-w2v_def variants (compared in both cases to 0 (perfect) when the same outputs are used). BLEU CV* is 2.154 for the NTS_def variants (compared to 0.838 for same outputs but different evaluation scripts, as above), and 6.598 for the NTS-w2v_def variants (compared to 1.314 for same outputs but different evaluation scripts). These differences arise simply from running the system in different environments.

The overall higher (worse) CV* values for NTS-w2v_def variants (compared to NTS_def) are likely to be partly due to the NTS models using one third party tool (openNMT), and the NTS-w2v models using two (openNMT and word2vec), i.e. the latter are more susceptible to changes in dependencies.

Object	Measurand	Measured quantity value							Sample size	mean	stdev	stdev 95% CI	CV* ↓
		Nisioi et al.	Cooper & Shardlow		this paper								
		outputs 1	outputs 1	outputs 2	outputs 1		outputs 3						
		s1 / b1	s1 / b2	s1 / b2	s1 / b3	s1 / b4	s1 / b3	s1 / b4					
NTS_def	BLEU	84.51	84.50	87.46	85.60	84.20	86.61	86.20	7	85.58	1.29	[0.45, 2.13]	1.562
	SARI	30.65	30.65	29.13	30.65		29.96		5	30.21	0.72	[0.095, 1.34]	2.487
NTS-w2v_def	BLEU	87.50	–	80.75	89.36	88.10	89.64	88.80	6	87.36	3.502	[0.92, 6.08]	4.176
	SARI	31.11	–	30.28	31.11		29.12		4	30.41	1.02	[-0.11, 2.15]	3.572

Table 4: Precision (CV*) and component measures (mean, standard deviation, standard deviation confidence intervals) for measured quantity values obtained in multiple measurements of the two **NTS systems**. Outputs 1 = test set outputs as generated by Nisioi et al. (2017); outputs 2 = test set outputs regenerated by Cooper and Shardlow (2020); outputs 3 = test set outputs regenerated by the present authors. s1 = SARI script (always the same); b1 = Nisioi et al.'s BLEU script, run by Nisioi et al.; b2 = Nisioi et al.'s BLEU script, run by Cooper & Shardlow; b3 = Nisioi et al.'s BLEU script with different version of NLTK tokeniser (see in text), run by the present authors; b4 = SacreBLEU (Xu et al., 2016), run by the present authors.

Object	Measurand	*Object conditions*		*Measurement method conditions*		*Measurement procedure conditions*			Measured quantity value	*CV**
		Code by	*Comp./trained by*	*Method*	*Implem. by*	*Procedure*	*Test set*	*Performed by*		
NTS_def	BLEU	Nisioi et al.	Nisioi et al.	bleu(o,t)	Nisioi et al.	OTE	Nisioi et al.	Nisioi et al.	**84.51**	*1.562*
		Nisioi et al.	Nisioi et al.	bleu(o,t)	Nisioi et al.	OTE	Nisioi et al.	Coop. & Shard.	84.50	
		Nisioi et al.	Nisioi et al.	bleu(o,t)	≈Nisioi et al.	OTE	Nisioi et al.	this paper	85.60	
		Nisioi et al.	Nisioi et al.	bleu(o,t)	SacreBLEU	OTE	Nisioi et al.	this paper	84.20	
		Nisioi et al.	Coop. & Shard.	bleu(o,t)	Nisioi et al.	OTE	Nisioi et al.	Coop. & Shard.	**87.46**	
		Nisioi et al.	this paper	bleu(o,t)	≈Nisioi et al.	OTE	Nisioi et al.	this paper	**86.61**	
		Nisioi et al.	this paper	bleu(o,t)	SacreBLEU	OTE	Nisioi et al.	this paper	86.20	
	SARI	Nisioi et al.	Nisioi et al.	sari(o,s,t)	Nisioi et al.	OITE	Nisioi et al.	Nisioi et al.	***30.65***	*2.487*
		Nisioi et al.	Nisioi et al.	sari(o,s,t)	Nisioi et al.	OITE	Nisioi et al.	Coop. & Shard.	*30.65*	
		Nisioi et al.	Nisioi et al.	sari(o,s,t)	Nisioi et al.	OITE	Nisioi et al.	this paper	*30.65*	
		Nisioi et al.	Coop. & Shard.	sari(o,s,t)	Nisioi et al.	OITE	Nisioi et al.	Coop. & Shard.	**29.13**	
		Nisioi et al.	this paper	sari(o,s,t)	Nisioi et al.	OITE	Nisioi et al.	this paper	**29.96**	
NTS-w2v_def	BLEU	Nisioi et al.	Nisioi et al.	bleu(o,t)	Nisioi et al.	OTE	Nisioi et al.	Nisioi et al.	**87.50**	*4.176*
		Nisioi et al.	Nisioi et al.	bleu(o,t)	≈Nisioi et al.	OTE	Nisioi et al.	this paper	89.36	
		Nisioi et al.	Nisioi et al.	bleu(o,t)	SacreBLEU	OTE	Nisioi et al.	this paper	88.10	
		Nisioi et al.	Coop. & Shard.	bleu(o,t)	Nisioi et al.	OTE	Nisioi et al.	Coop. & Shard.	**80.75**	
		Nisioi et al.	this paper	bleu(o,t)	≈Nisioi et al.	OTE	Nisioi et al.	this paper	**89.64**	
		Nisioi et al.	this paper	bleu(o,t)	SacreBLEU	OTE	Nisioi et al.	this paper	88.80	
	SARI	Nisioi et al.	Nisioi et al.	sari(o,s,t)	Nisioi et al.	OITE	Nisioi et al.	Nisioi et al.	***31.11***	*3.572*
		Nisioi et al.	Nisioi et al.	sari(o,s,t)	Nisioi et al.	OITE	Nisioi et al.	this paper	*31.11*	
		Nisioi et al.	Coop. & Shard.	sari(o,s,t)	Nisioi et al.	OITE	Nisioi et al.	Coop. & Shard.	**30.28**	
		Nisioi et al.	this paper	sari(o,s,t)	Nisioi et al.	OITE	Nisioi et al.	this paper	**29.12**	

Table 5: Conditions of measurement for each measurement carried out for the **NTS systems**. OTE = outputs vs. targets evaluation, OITE = outputs vs. inputs and targets evaluation. Shaded cells: evaluation of the same system outputs, i.e. the reproductions did not regenerate outputs. Bold: evaluation of (potentially) different system outputs, i.e. the reproductions did regenerate outputs.

4.2 QRA for PASS system

The PASS system, developed by van der Lee et al. (2017), generates football match reports from the perspective of each of the competing teams. The original study evaluated the system for Clarity, Fluency and Stance Identifiability in an evaluation with 20 evaluators and a test set of 10 output pairs. The evaluation was repeated with a slightly different evaluation interface and a different cohort of evaluators by Mille et al. (2021). Table 2 shows the results from the original and reproduction evaluations (columns 3 and 4), where the Clarity and Fluency results are the mean scores from 7-point agreement scales, and Identifiability results are the percentage of times the evaluators correctly guessed the team whose supporters a report was written for. Columns 6–9 show the corresponding sample size (number of reproductions plus original study), mean, standard deviation (stdev), the confidence interval (CI) for the standard deviation, and CV*, all calculated on the shifted scores (see Section 3.2).

Table 3 shows the values (here, paper identifiers) for the nine conditions of measurement introduced in Section 3.3, for each of the six individual measurements (three evaluation measures times two studies). Note that both object conditions and the test set condition are the same, because Mille et al. used the system outputs shared by van der Lee et al. The values for the *Implemented by*, *Procedure* and *Performed by* conditions reflect the differences in the two evaluations in design, evaluator cohorts, and the teams that performed them.

The scores vary to different degrees for the three measurands, with CV* lowest (reproducibility best) for Stance Identifiability, and highest (worst) for Fluency. These CV* results are likely to reflect that evaluators agreed more on Clarity than Fluency. Moreover, the binary stance identification assessment has better reproducibility than the other two criteria which are assessed on 7-point rating scales.

4.3 QRA for essay scoring system variants

The 11 multilingual essay scoring system variants reported by Vajjala and Rama (2018) were evaluated by weighted F1 (wF1) score. Table 6 shows wF1 scores for the 11 multilingual system variants from each of the five papers, alongside the 11 corresponding CV* values. Table 7 in the appendix shows the corresponding conditions of measurement. The baseline classifier (mult-base) uses document length (number of words) as its only feature. For the other variants, +/- indicates that the multilingual classifier was / was not given information about which language the input was in; the mult-word variants use word n-grams only; mult-word uses POS (part of speech) tag n-grams only; mult-dep uses n-grams over dependency relation, dependent POS, and head POS triples; mult-dom uses domain-specific linguistic features including document length, lexical richness and errors; mult-emb uses word and character embeddings. The mult-base and mult-dom models are logistic regressors, the others are random forests.

A very clear picture emerges: system variant pairs that differ only in whether they do or do not use language information have very similar CV scores. For example, mult-POS$^-$ (POS n-grams without language information) and mult-POS$^+$ (POS n-grams with language information) both have a very good degree of wF1-reproducibility, their CV* being 3.818 and 3.808 respectively; mult-word$^-$ (word n-grams without language information) and mult-word$^+$ (word n-grams with language information) have notably higher CV*, around 10. This tendency holds for all such pairs, indicating that using language information makes next to no difference to reproducibility. Moreover, the mult-dom and mult-emb variants all have similar CV*.[9]

The indication is that the syntactic information is obtained/used in a way that is particularly reproducible, whereas the domain-specific information and the embeddings are obtained/used in a way that is particularly hard to reproduce. Overall, the random forest models using syntactic features have the best reproducibility; the logistic regressors using domain-specific features have the worst.

5 Discussion

Quantified reproducibility assessment (QRA) enables assessment of the degree of reproducibility of evaluation results for any given system and evaluation measure in a way that is scale-invariant[10] and comparable across different QRAs, for reproductions involving either the same or different original studies. Moreover, formally capturing (dis)similarities between systems and evaluation designs enables reproducibility to be assessed relative to such (dis)similarities. In combination, a set of results from QRA tests for the same system and evaluation measure can provide pointers to which aspects of the system and evaluation might be associated with low reproducibility. E.g. for the wF1 evaluations of the essay scoring systems above, it is clear that variations in reproducibility are associated at least in part with the different features used by systems.

It might be expected that the reproducibility of human-assessed evaluations is generally worse than metric-assessed. Our study revealed a more mixed picture. As expected, the Fluency and Clarity evaluations of the PASS system were among those with highest CV*, and the BLEU and SARI evaluation of the NTS systems and wF1 evaluation of the mult-POS and mult-dep systems were among those with lowest CV*. However, human-assessed Stance Identifiability of PASS was among the most reproducible, and metric-assessed wF1 of mult-base, mult-dom and mult-emb were among the worst.

In this paper, our focus has been QRA testing of existing research results. However, ideally, QRA would be built into new method development from the outset, where at first reporting, a detailed stan-

[9] The high CV* for the baseline system may be due to an issue wiith the evaluation code (macro-F1 instead of weighted-F1), as reported by Bestgen (Section 3.2, first paragraph), Caines and Buttery (Section 2.5, one before last paragraph) and Huber and Çöltekin (Section 3.2, second paragraph).

[10] If evaluation scores are multiplied by a common factor, CV* does not change.

Object	Meas-urand	Measured quantity value								Sample size	mean	stdev	stdev 95% CI	CV* ↓
		Vajjala & Rama	Huber & Coltekin	Arhiliuc et al.	Bestgen			Caines & Buttery						
		seed 1	seed 2	seed ?	seed 1		seed 2	seed 1	seed ?					
		e1 / i1	e2 / i2	e3 / i1	e4 / i1	e5 / i1	e5 / i3	e6 / i1	e7 / i4					
mult-base	wF1	0.428	0.493	0.426	0.574	0.579	0.590	0.574	0.600	8	0.533	0.08	[0.03, 0.12]	14.633
mult-word$^-$	wF1	0.721	0.603	0.605	0.606	0.720	0.732	0.606	0.740	8	0.667	0.07	[0.03, 0.11]	10.609
mult-word$^+$	wF1	0.719	0.604	0.607	0.607	0.723	0.733	0.607	0.736	8	0.667	0.07	[0.03, 0.11]	10.440
mult-POS$^-$	wF1	0.726	0.681	0.680	0.680	0.722	0.728	0.680	0.732	8	0.704	0.03	[0.01, 0.04]	3.818
mult-POS$^+$	wF1	0.724	0.680	0.680	0.681	0.725	0.729	0.681	0.731	8	0.704	0.03	[0.01, 0.04]	3.808
mult-dep$^-$	wF1	0.703	0.660	0.650	0.651	0.699	0.711	0.651	0.710	8	0.679	0.03	[0.01, 0.05]	4.500
mult-dep$^+$	wF1	0.693	0.661	0.652	0.653	0.699	0.712	0.653	0.716	8	0.68	0.03	[0.01, 0.05]	4.387
mult-dom$^-$	wF1	0.449	0.600	0.433	0.597	0.635	0.646	0.597	0.698	8	0.582	0.1	[0.04, 0.15]	17.147
mult-dom$^+$	wF1	0.471	0.647	0.447	0.647	0.696	0.711	0.647	0.726	8	0.624	0.11	[0.05, 0.18]	18.248
mult-emb$^-$	wF1	0.693	0.658	0.683	0.668	0.692	0.689	0.659	0.391	8	0.642	0.11	[0.04, 0.17]	17.033
mult-emb$^+$	wF1	0.689	0.662	0.681	0.659	0.681	0.684	0.657	0.401	8	0.639	0.1	[0.04, 0.16]	16.226

Table 6: Precision (CV*) and component measures (mean, standard deviation, standard deviation confidence intervals) for measured quantity values obtained in multiple measurements of the **essay scoring systems**. Seed i = different approaches to random seeding and cross-validation; ei = different compile/run-time environments; ii = different test data sets and/or cross-validation folds.

dardised set of conditions of measurement is specified, and repeatability tests (where all conditions are identical except for the team conducting the tests, see Section 3.2) are performed to determine baseline reproducibility. Such repeatability QRA would provide quality assurance for new methods as well as important pointers for future reproductions regarding what degree of reproducibility to expect for given (types of) methods.

If this is not possible, post-hoc reproducibility QRA (where there are differences in conditions of measurement values) is performed instead. If this yields high (poor) CV*, one way to proceed is to minimise differences in conditions of measurement between the studies and observe the effect on CV*, changing aspects of system and evaluation design and adding further conditions of measurement if need be. For human evaluation in particular, persistently high CV* would indicate a problem with the method itself.

6 Conclusion

We have described an approach to quantified reproducibility assessment (QRA) based on concepts and definitions from metrology, and tested it on 18 system and evaluation measure combinations involving diverse NLP tasks and types of evaluation.

QRA produces a single score that quantifies the degree of reproducibility of a given system and evaluation measure, on the basis of the scores from, and differences between, multiple reproductions of the same original study. We found that the approach facilitates insights into sources of variation between reproductions, produces results that are comparable across different reproducibility assessments, and provides pointers about what needs to be changed in system and/or evaluation design to improve reproducibility.

A recent survey (Belz et al., 2021a) found that just 14% of the 513 original/reproduction score pairs analysed were exactly the same. Judging the remainder simply ‘not reproduced’ is of limited usefulness, as some are much closer to being the same than others. At the same time, assessments of whether the same conclusions can be drawn on the basis of different scores involve subjective judgments and are prone to disagreement among assessors. Quantifying the closeness of results as in QRA, and, over time, establishing expected levels of closeness, seems a better way forward.

Acknowledgements

We are grateful to the anonymous reviewers and area chairs for their exceptionally detailed and helpful feedback.

Popović’s work on this s study was funded by the ADAPT SFI Centre for Digital Media Technology which is funded by Science Foundation Ireland through the SFI Research Centres Programme, and co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106. Mille’s work was supported by the European Commission under the H2020 program contract numbers 786731, 825079, 870930 and 952133.

References

SE Ahmed. 1995. A pooling methodology for coefficient of variation. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 57–75.

Cristina Arhiliuc, Jelena Mitrović, and Michael Granitzer. 2020. Language proficiency scoring. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5624–5630, Marseille, France. European Language Resources Association.

Association for Computing Machinery. 2020. Artifact review and badging Version 1.1. Accessed August 24, 2020.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021a. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393.

Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021b. The reprogen shared task on reproducibility of human evaluations in NLG: Overview and results. In *The 14th International Conference on Natural Language Generation*.

Yves Bestgen. 2020. Reproducing monolingual, multilingual and cross-lingual CEFR predictions. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5595–5602, Marseille, France. European Language Resources Association.

António Branco, Nicoletta Calzolari, Piek Vossen, Gertjan Van Noord, Dieter van Uytvanck, João Silva, Luís Gomes, André Moreira, and Willem Elbers. 2020. A shared task of a new, collaborative type to foster reproducibility: A first exercise in the area of language science and technology with REPROLANG2020. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5539–5545, Marseille, France. European Language Resources Association.

Andrew Caines and Paula Buttery. 2020. REPROLANG 2020: Automatic proficiency scoring of Czech, English, German, Italian, and Spanish learner essays. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5614–5623, Marseille, France. European Language Resources Association.

Michael Cooper and Matthew Shardlow. 2020. CombiNMT: An exploration into neural text simplification models. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5588–5594, Marseille, France. European Language Resources Association.

Chris Drummond. 2009. Replicability is not reproducibility: nor is it good science. Presented at 4th Workshop on Evaluation Methods for Machine Learning held at ICML'09.

Eva Huber and Çağrı Çöltekin. 2020. Reproduction and replication: A case study with automatic essay scoring. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5603–5613, Marseille, France. European Language Resources Association.

JCGM. 2012. International vocabulary of metrology: Basic and general concepts and associated terms (VIM). Joint Committee for Guides in Metrology, `https://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf`.

Simon Mille, Thiago Castro Ferreira, Anya Belz, and Brian Davis. 2021. Another PASS: A reproduction study of the human evaluation of a football report generation system. In *Proceedings of the 14th International Conference on Natural Language Generation (INLG 2021)*.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Ted Pedersen. 2008. Empiricism is not a matter of faith. *Computational Linguistics*, 34(3):465–470.

Joelle Pineau. 2020. The machine learning reproducibility checklist v2.0.

Maja Popović and Anya Belz. 2021. A reproduction study of an annotation-based human evaluation of MT outputs. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 293–300, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores. *WMT 2018*, page 186.

Calyampudi Radhakrishna Rao. 1973. *Linear statistical inference and its applications*. Wiley.

Nicolas P. Rougier, Konrad Hinsen, Frédéric Alexandre, Thomas Arildsen, Lorena A Barba, Fabien CY Benureau, C Titus Brown, Pierre De Buyl, Ozan Caglayan, Andrew P Davison, et al. 2017. Sustainable computational science: The ReScience initiative. *PeerJ Computer Science*, 3:e142.

Patrick D. Schloss. 2018. Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *MBio*, 9(3).

Anastasia Shimorina and Anya Belz. 2021. The human evaluation datasheet 1.0: A template for recording details of human evaluation experiments in NLP. *arXiv preprint arXiv:3910940*.

R.R. Sokal and F.J. Rohlf. 1971. *Biometry: The Principles and Practice of Statistics in Biological Research*. WH Freeman.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153, New Orleans, Louisiana. Association for Computational Linguistics.

Chris van der Lee, Emiel Krahmer, and Sander Wubben. 2017. PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 95–104.

Kirstie Whitaker. 2017. The MT Reproducibility Checklist. https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf.

Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics*, 44(4):641–649.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

A Conditions of Measurement for the Essay Scoring Systems

Table 7 shows the conditions of measurement for each of the 88 individual measurements for the Essay Scoring Systems.

Object	Measurand	*Object conditions*		*Measurement method conditions*		*Measurement procedure conditions*			Measured quantity value	*CV**
		Code by	*Comp./trained by*	*Method*	*Implem. by*	*Procedure*	*Test set*	*Performed by*		
mult-base	wF1	Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Va.& Ra.	0.428	*14.633*
		Va.& Ra.	Huber & Coltekin	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Huber & Coltekin	0.493	
		Va.& Ra.	Arhiliuc et al.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Arhiliuc et al.	0.426	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.574	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.579	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	≈Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.590	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Cai. & But.	0.574	
		Cai. & But.	Cai. & But.	wF1(o,t)	Cai. & But.	OTE	Va.&Ra.	Cai. & But.	0.600	
mult-word^{-}	wF1	Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Va.& Ra.	0.721	*10.609*
		Va.& Ra.	Huber & Coltekin	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Huber & Coltekin	0.603	
		Va.& Ra.	Arhiliuc et al.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Arhiliuc et al.	0.605	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.606	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.720	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	≈Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.732	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Cai. & But.	0.606	
		Cai. & But.	Cai. & But.	wF1(o,t)	Cai. & But.	OTE	Va.&Ra.	Cai. & But.	0.740	
mult-word^{+}	wF1	Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Va.& Ra.	0.719	*10.44*
		Va.& Ra.	Huber & Coltekin	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Huber & Coltekin	0.604	
		Va.& Ra.	Arhiliuc et al.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Arhiliuc et al.	0.607	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.607	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.723	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	≈Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.733	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Cai. & But.	0.607	
		Cai. & But.	Cai. & But.	wF1(o,t)	Cai. & But.	OTE	Va.&Ra.	Cai. & But.	0.736	
mult-POS^{-}	wF1	Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Va.& Ra.	0.726	*3.818*
		Va.& Ra.	Huber & Coltekin	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Huber & Coltekin	0.681	
		Va.& Ra.	Arhiliuc et al.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Arhiliuc et al.	0.680	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.680	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.722	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	≈Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.728	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Cai. & But.	0.680	
		Cai. & But.	Cai. & But.	wF1(o,t)	Cai. & But.	OTE	Va.&Ra.	Cai. & But.	0.732	
mult-POS^{+}	wF1	Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Va.& Ra.	0.724	*3.808*
		Va.& Ra.	Huber & Coltekin	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Huber & Coltekin	0.680	
		Va.& Ra.	Arhiliuc et al.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Arhiliuc et al.	0.680	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.681	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.725	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	≈Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.729	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Cai. & But.	0.681	
		Cai. & But.	Cai. & But.	wF1(o,t)	Cai. & But.	OTE	Va.&Ra.	Cai. & But.	0.731	
mult-dep^{-}	wF1	Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Va.& Ra.	0.703	*4.5*
		Va.& Ra.	Huber & Coltekin	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Huber & Coltekin	0.660	
		Va.& Ra.	Arhiliuc et al.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Arhiliuc et al.	0.650	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.651	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.699	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	≈Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.711	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Cai. & But.	0.651	
		Cai. & But.	Cai. & But.	wF1(o,t)	Cai. & But.	OTE	Va.&Ra.	Cai. & But.	0.710	
mult-dep^{+}	wF1	Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Va.& Ra.	0.693	*4.387*
		Va.& Ra.	Huber & Coltekin	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Huber & Coltekin	0.661	
		Va.& Ra.	Arhiliuc et al.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Arhiliuc et al.	0.652	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.653	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.699	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	≈Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.712	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Cai. & But.	0.653	
		Cai. & But.	Cai. & But.	wF1(o,t)	Cai. & But.	OTE	Va.&Ra.	Cai. & But.	0.716	

Table continued on next page.

Object	Measurand	*Object conditions*		*Measurement method conditions*		*Measurement procedure conditions*			Measured quantity value	*CV**
		Code by	*Comp./trained by*	*Method*	*Implem. by*	*Procedure*	*Test set*	*Performed by*		
mult-dom^{-}	wF1	Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Va.& Ra.	0.449	*17.147*
		Va.& Ra.	Huber & Coltekin	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Huber & Coltekin	0.600	
		Va.& Ra.	Arhiliuc et al.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Arhiliuc et al.	0.433	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.597	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.635	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	≈Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.646	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Cai. & But.	0.597	
		Cai. & But.	Cai. & But.	wF1(o,t)	Cai. & But.	OTE	Va.&Ra.	Cai. & But.	0.698	
mult-dom^{+}	wF1	Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Va.& Ra.	0.471	*18.248*
		Va.& Ra.	Huber & Coltekin	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Huber & Coltekin	0.647	
		Va.& Ra.	Arhiliuc et al..	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Arhiliuc et al.	0.447	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.647	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.696	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	≈Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.711	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Cai. & But.	0.647	
		Cai. & But.	Cai. & But.	wF1(o,t)	Cai. & But.	OTE	Va.&Ra.	Cai. & But.	0.726	
mult-emb^{-}	wF1	Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Va.& Ra.	0.693	*17.033*
		Va.& Ra.	Huber & Coltekin	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Huber & Coltekin	0.658	
		Va.& Ra.	Arhiliuc et al.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Arhiliuc et al.	0.683	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.668	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.692	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	≈Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.689	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Cai. & But.	0.659	
		Cai. & But.	Cai. & But.	wF1(o,t)	Cai. & But.	OTE	Va.&Ra.	Cai. & But.	0.391	
mult-emb^{+}	wF1	Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Va.& Ra.	0.689	*16.226*
		Va.& Ra.	Huber & Coltekin	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Huber & Coltekin	0.662	
		Va.& Ra.	Arhiliuc et al.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Arhiliuc et al.	0.681	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.659	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.681	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	≈Va.& Ra.	OTE	Va.& Ra.	Bestgen	0.684	
		Va.& Ra.	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra.	Cai. & But.	0.657	
		Cai. & But.	Cai. & But.	wF1(o,t)	Cai. & But.	OTE	Va.&Ra.	Cai. & But.	0.401	

Table 7: Conditions of measurement for each measurement carried out for the multilingual **essay scoring systems**. OTE = outputs vs.targets evaluation.