A Metrological Perspective on Reproducibility in NLP*

Anya Belz
ADAPT Research Centre,
Dublin City University
University of Aberdeen
anya.belz@adaptcentre.ie

Reproducibility has become an increasingly debated topic in NLP and ML over recent years, but so far, no commonly accepted definitions of even basic terms or concepts have emerged. The range of different definitions proposed within NLP/ML not only do not agree with each other, they are also not aligned with standard scientific definitions. This article examines the standard definitions of repeatability and reproducibility provided by the meta-science of metrology, and explores what they imply in terms of how to assess reproducibility, and what adopting them would mean for reproducibility assessment in NLP/ML. It turns out the standard definitions lead directly to a method for assessing reproducibility in quantified terms that renders results from reproduction studies comparable across multiple reproductions of the same original study, as well as reproductions of different original studies. The article considers where this method sits in relation to other aspects of NLP work one might wish to assess in the context of reproducibility.

1. Introduction

Reproducibility of results is coming under increasing scrutiny in the machine learning (ML) and natural language processing (NLP) fields, against the background of a perceived reproducibility crisis in science more widely (Baker 2016), and NLP/ML specifically (Mieskes et al. 2019). There have been workshops and checklist initiatives, conferences promoting reproducibility via calls, chairs' blogs and special themes, and the first shared tasks, including REPROLANG'20 (Branco et al. 2020) and ReproGen'21 (Belz et al. 2021b).

Despite this growing body of research on reproducibility, it has been observed that no standard terms and definitions have emerged (Cohen et al. 2018). For the Association of Computing Machinery (ACM 2020), *results* have been *reproduced* if obtained in a

* This article makes the general case for (i) adopting general scientific definitions of reproducibility and repeatability, and (ii) computing precision as the quantified measure of the degree of reproducibility of scores, as in other scientific disciplines. The QRA approach itself is described and discussed in more detail in Belz (2021), and tested on a variety of reproduction scenarios in Belz, Popovic, and Mille (2022). The code for computing CV* and several example score sets from the above papers can be found here: github.com/asbelz/coeff-var.

Submission received: 28 September 2021; revised version received: 27 April 2022; accepted for publication: 2 June 2022.

https://doi.org/10.1162/coli_a_00448

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. For Drummond (2009), *replicability* is defined as an experiment being able to be re-run exactly, whereas *reproducibility* is the ability to obtain the same result by different means. For Rougier et al. (2017), "*[r]eproducing* the result of a computation means running the same software on the same input data and obtaining the same results. [...]. *Replicating* a published result means writing and then running new software based on the description of a computational model or method." Wieling, Rawee, and van Noord (2018) tie *reproducibility* to "the same data and methods," and Whitaker (2017), followed by Schloss (2018), tie definitions of *reproducibility, replicability, robustness*, and *generalizability* to different combinations of same vs. different data and code. For Cohen et al. (2018) *reproducibility* is "a property of the outcomes of an experiment: arriving—or not—at the same conclusions, findings, or values," whereas they "exclude issues related to the ability to repeat the experiments reported in a paper [which is] replicability or repeatability."

No two of the above definitions are fully in agreement with each other, and none are entirely aligned with the standard scientific definitions of repeatability and reproducibility used in other fields and codified in the International Vocabulary of Metrology (VIM) (JCGM 2012). In fact, the US National Information Standards Organization had to ask the ACM to "harmonize its terminology and definitions with those used in the broader scientific research community" (ACM 2020), which prompted the latter to switch its definitions of replicability and reproducibility, without however achieving full alignment with the standard definitions of the two terms.

The remainder of this article examines the VIM definitions and the kind of approach to assessing reproducibility their adoption for NLP/ML would result in. Section 2 presents and discusses the VIM definitions (JCGM 2012), Section 3 lays out what reproducibility assessment looks like if wholly based on them, including practical steps and an example application in NLP. Section 4 discusses the limits of the proposed approach and a variety of questions that have come up in discussions and reviews of this work. The article concludes (Section 5) with a consideration of what would need to change for Quantified Reproducibility Assessment (QRA) to become part of NLP workflows, and what incentive the field might have to accept the implied overhead.

2. Definitions from VIM and Implications for Reproducibility Assessment

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.[1]

1 In physical measurements objects can change simply as a function 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, repeatability and reproducibility are properties not of objects, scores, results, or conclusions, but of measurements M that are parameterized by measurand m, object O, time t, and conditions of measurement C and return a measured quantity value v. They are defined as measurement precision, that is, quantified by calculating the precision of a set of measured quantity values, relative to a set of conditions of measurement, which have to be known and specified for assessment of repeatability and reproducibility to be meaningful.

These definitions map directly to the following formal definitions of repeatability and reproducibility. First, repeatability R^0 (where all conditions of measurement are fixed):

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

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. Reproducibility R is defined in the same way except that condition values differ for at least one condition in the M_i (hence C_i is here subscripted):

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

VIM does not tell us how to compute precision. Below, the unbiased coefficient of variation is used, but other measures are possible. The members of each set of conditions can be construed as attribute/value pairs each consisting of a name and a value. VIM does not tell us which exact set of measurement conditions to use; these depend on measurement type.

3. Assessing Reproduction Results in Practice

Generally, when carrying out a reproduction study of some original work in ML/NLP, we consider the system and how it was created and trained, the scores that were obtained for it, and the claims made on the basis of the scores. In existing work, the corresponding questions that are addressed in assessing the results of reproduction are of three broad types: how easily can the system be recreated; how similar are the scores; and can the same claims be supported. Only the second of these is about reproducibility in the general scientific sense, and only is it answerable from metrology. In order to have a complete approach to answering it in practice we need three more components in addition to the definitions from Section 2: (i) a method for computing precision; (ii) specified conditions of measurement; and (iii) a procedure for carrying out reproducibility assessments. Each is addressed in turn below, followed by an example application.

Computing Precision. In other fields, measurement precision is typically reported in terms of the coefficient of variation (CV) defined as the standard deviation over the mean (of a sample of measured quantity values). Other values reported can include: mean along side standard deviation with 95% confidence intervals, coefficient of variation, and percentage of values within n standard deviations. CV serves well as the "headline" result, because it is a general measure, not in the unit of the measurements (unlike mean and standard deviation), providing a quantification of degree of reproducibility that is comparable across studies (Ahmed 1995, page 57). This also holds for percentage within n standard deviations but the latter is less recognized and intuitive.

In reproduction studies in NLP/ML, sample sizes tend to be very small. 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 s^* approximated on the basis of the standard error of the unbiased sample variance $\mathrm{se}(s^2)$ as $\mathrm{se}_{s^2}(s^*) \approx \frac{1}{2\sigma}\mathrm{se}(s^2)$ (Rao 1973). Assuming that measured quantity values are normally distributed, we calculate the standard error of the sample variance in the usual way: $\mathrm{se}(s^2) = \sqrt{\frac{2\sigma^4}{n-1}}$. Finally, we also use a small sample correction (Sokal and Rohlf 1971) for the coefficient of variation, which gives us the unbiased coefficient of variation CV^*, defined as follows:[2]

$$\mathrm{CV}^* = \left(1 + \frac{1}{4n}\right)\frac{s^*}{|m|} \tag{3}$$

where s^* is the unbiased standard deviation, and m the sample mean.

Before applying CV^* to values on scales that do not start at 0 (in NLP this happens mostly in human evaluations) values need to be shifted to start at 0 to ensure comparability.[3]

Conditions of Measurement. Individual conditions are specified by name and value, and their role is to capture those attributes of a measurement where different values may cause differences in measured quantity values. For repeatability assessment, conditions need to capture all sources of variation in results (although not necessarily each with a separate condition). It so happens that much of the reproducibility work in ML and NLP has so far focused on what standard conditions of measurement (information about system, data, dependencies, computing environment, etc.) for metric measurements need to be specified in order to enable repeatability assessment, even if it hasn't been couched in these terms.

Reproducibility checklists such as those provided by Pineau (2020) and the ACL[4] for metric evaluation, and evaluation datasheets like the one proposed by Shimorina and Belz (2021) for human evaluation, are lists of types of information (attributes), for which authors are asked to provide information (values), that can directly be construed as conditions of measurement. In the example in the following section, we use the following conditions of measurement, based on the above checklists and described in more detail in Belz, Popovic, and Mille (2022):

1. *Object conditions*: (a) System code, (b) Compile/training information.

2 Code and data are available here: `https://github.com/asbelz/coeff-var`.
3 Otherwise CV^* reflects differences solely due to different minimum values.
4 `https://2021.aclweb.org/calls/reproducibility-checklist/`.

2. *Measurement method conditions:*[5] (a) Measurement method specification, (b) Measurement method implementation.
3. *Measurement procedure conditions:*[6] (a) Measurement procedure, (b) Test set, (c) Performed by.

The *names* of the conditions of measurement used in this paper are as given above. The *values* for each condition characterize how measurements differ in respect to the condition; in reporting results from QRA tests below, we use paper and method identifiers as shorthand for distinct condition values (full details in each case being available from the referenced papers). The intention here is not to propose a definitive set of conditions for general use; that is beyond the scope of this article, and at any rate should evolve by consensus over time.

Reproducibility Assessment Procedure. We now have all the components needed for quantified reproducibility assessment, but how do we go about carrying it out in practice? From metrology we know we need to compute the precision of two or more measured quantity values obtained in measurements of the same object and measurand with specified conditions of measurement, identical in the case of repeatability and differing in the case of reproducibility. There are two possible starting positions: (a) the original and (some) reproduction studies have been carried out; and (b) the original study has been carried out, but no reproductions.

In the case of *a*, the conditions of measurement that can be used are limited to the information that can be gleaned from publications, code/data repositories, and the authors. This rules out repeatability assessment in most cases. In such cases, reproducibility assessment involves the steps indicated in the bottom half of Figure 1, termed 1-Phase QRA. The example QRA below is such a case. If the set of measurements includes subsets of measurements that have more conditions in common, then it makes sense to report *R* for these subsets separately.

In the case of *b*, repeatability assessment can be carried out, and arguably should be carried out before any reproducibility assessment. For a true estimate of the variation resulting from given differences in condition values, the baseline variation, present when all condition values are the same, needs to be known. Repeatability assessment should therefore be carried out prior to reproducibility assessment. For example, if the coefficient of variation is x_{C^0} under identical conditions C^0, and x_{C_i} under varied conditions C_i, then it is the difference between x_{C^0} and x_{C_i} that estimates the effect of varying the conditions. The steps in this 2-phase reproducibility assessment are shown at the top of Figure 1.

Example Application.[7] As an example of how QRA can be used in practice we apply the approach outlined above, in its 1-phase version with conditions of measurement and CV* as given above, to a set of reproductions of the essay scoring system evaluations reported by Vajjala and Rama (2018), which were carried out as part of REPROLANG (Branco et al. 2020). More particularly, we look at five of the 11 multilingual essay scoring system variants and the weighted F1 scores (wF1) computed for them. Table 1

5 For definition of 'measurement method', see VIM 2.5.
6 For definition of 'measurement procedure', see VIM 2.6.
7 We report four example applications, including human evaluations, in full detail elsewhere (Belz, Popovic, and Mille 2022), including the other six multilingual system variants from the Vajjala and Rama (2018) reproductions below.

2-PHASE QUANTIFIED REPRODUCIBILITY ASSESSMENT

REPEATABILITY PHASE

1. Select measurement to be assessed, identify shared object and measurand.
2. Select initial set of repeatability conditions of measurement C^0, specify value for each condition.
3. Perform $n \geq 2$ reproduction measurements to yield measured quantity values $v_1^0, v_2^0, \ldots v_n^0$.
4. Compute precision for $v_1^0, v_2^0, \ldots v_n^0$, giving repeatability score R^0.
5. Unless precision is as small as desired, identify additional conditions that had different values in some of the reproduction measurements, and add them to the set of measurement conditions, also updating the measurements to ensure same values for the new conditions. Repeat Steps 3–5.

REPRODUCIBILITY PHASE

6. From the final set of repeatability conditions, select the conditions to vary, and specify the different values to test.
7. For each combination of differing condition values:
   (a) Carry out n reproduction tests, yielding measured quantity values $v_1, v_2, \ldots v_n$.
   (b) Compute precision for $v_1, v_2, \ldots v_n$, giving reproducibility score R.

Report all resulting R scores, alongside baseline R^0 score.

1-PHASE QUANTIFIED REPRODUCIBILITY ASSESSMENT

1. For set of n measurements to be assessed, identify object and measurand.
2. Identify all conditions of measurement C for which information is available for all measurements M_i, and specify condition values for each measurement.
3. Gather the n measured quantity values $v_1, v_2, \ldots v_n$.
4. Compute precision of $v_1, v_2, \ldots v_n$, giving reproducibility score R.

Report resulting R score.

Figure 1
Top: Steps in 2-Phase QRA with baseline repeatability assessment (omit Step 5 for standard conditions of measurement). Bottom: Steps in 1-Phase QRA where baseline repeatability assessment is impossible).

shows object (system variant) and measurand (wF1) in the first two columns. 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; mult-dep uses n-grams over dependency relation, dependent POS, and head POS triples; and mult-emb uses word and character embeddings. The mult-base model is a logistic regressor, the others are random forests.

As can be see from the *Performed by* column and the measured quantity values (wF1 scores) in the second-to-last column, for each system variant, we have the original

Table 1
QRA results for five of the multilingual essay scoring system variants reported by Vajjala and Rama (2018). Reproductions by Huber and Çöltekin (2020); Arhiliuc, Mitrović, and Granitzer (2020); Bestgen (2020); Caines and Buttery (2020). OTE = outputs vs. targets evaluation; s*i* = different random seeding method; env*i* = different compile/run-time environments; d*i* = different test data splits.

Object	Meas-	*Object conditions*		*Measurement method conditions*		*Measurement procedure conditions*			Measured quantity	*CV**
	urand	*Code by*	*Compiled/ trained by*	*Method*	*Implemented by*	*Proc- edure*	*Test set*	*Performed by*	value	
mult-base	wF1	Va.& Ra./s1	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Va.& Ra.	0.428	*14.63*
		Va.& Ra./s2	Hub.&Colt.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d2	Hub.&Colt.	0.493	
		Va.& Ra./s?	Arhiliuc&al.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Arhiliuc&al.	0.426	
		Va.& Ra./s1	Va.& Ra./env1	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Bestgen	0.574	
		Va.& Ra./s1	Va.& Ra./env2	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Bestgen	0.579	
		Va.& Ra./s2	Va.& Ra./env2	wF1(o,t)	≈Va.& Ra.	OTE	Va.& Ra./d3	Bestgen	0.590	
		Va.& Ra./s1	Va.& Ra./env3	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Cai.&But.	0.574	
		Cai.&But.	Cai.&But.	wF1(o,t)	Cai.&But.	OTE	Va.&Ra./d4	Cai.&But.	0.600	
mult-dep$^-$	wF1	Va.& Ra./s1	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Va.& Ra.	0.703	*4.5*
		Va.& Ra./s2	Hub.&Colt.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d2	Hub.&Colt.	0.660	
		Va.& Ra./s?	Arhiliuc&al.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Arhiliuc&al.	0.650	
		Va.& Ra./s1	Va.& Ra./env1	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Bestgen	0.651	
		Va.& Ra./s1	Va.& Ra./env2	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Bestgen	0.699	
		Va.& Ra./s2	Va.& Ra./env2	wF1(o,t)	≈Va.& Ra.	OTE	Va.& Ra./d3	Bestgen	0.711	
		Va.& Ra./s1	Va.& Ra./env3	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Cai.&But.	0.651	
		Cai.&But.	Cai.&But.	wF1(o,t)	Cai.&But.	OTE	Va.&Ra./d4	Cai.&But.	0.710	
mult-dep$^+$	wF1	Va.& Ra./s1	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Va.& Ra.	0.693	*4.39*
		Va.& Ra./s2	Hub.&Colt.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d2	Hub.&Colt.	0.661	
		Va.& Ra./s?	Arhiliuc&al.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Arhiliuc&al.	0.652	
		Va.& Ra./s1	Va.& Ra./env1	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Bestgen	0.653	
		Va.& Ra./s1	Va.& Ra./env2	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Bestgen	0.699	
		Va.& Ra./s2	Va.& Ra./env2	wF1(o,t)	≈Va.& Ra.	OTE	Va.& Ra./d3	Bestgen	0.712	
		Va.& Ra./s1	Va.& Ra./env3	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Cai.&But.	0.653	
		Cai.&But.	Cai.&But.	wF1(o,t)	Cai.&But.	OTE	Va.&Ra./d4	Cai.&But.	0.716	
mult-emb$^-$	wF1	Va.& Ra./s1	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Va.& Ra.	0.693	*17.03*
		Va.& Ra./s2	Hub.&Colt.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d2	Hub.&Colt.	0.658	
		Va.& Ra./s?	Arhiliuc&al.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Arhiliuc&al.	0.683	
		Va.& Ra./s1	Va.& Ra./env1	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Bestgen	0.668	
		Va.& Ra./s1	Va.& Ra./env2	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Bestgen	0.692	
		Va.& Ra./s2	Va.& Ra./env2	wF1(o,t)	≈Va.& Ra.	OTE	Va.& Ra./d3	Bestgen	0.689	
		Va.& Ra./s1	Va.& Ra./env3	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Cai.&But.	0.659	
		Cai.&But.	Cai.&But.	wF1(o,t)	Cai.&But.	OTE	Va.&Ra./d4	Cai.&But.	0.391	
mult-emb$^+$	wF1	Va.& Ra./s1	Va.& Ra.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Va.& Ra.	0.689	*16.23*
		Va.& Ra./s2	Hub.&Colt.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d2	Hub.&Colt.	0.662	
		Va.& Ra./s?	Arhiliuc&al.	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Arhiliuc&al.	0.681	
		Va.& Ra./s1	Va.& Ra./env1	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Bestgen	0.659	
		Va.& Ra./s1	Va.& Ra./env2	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Bestgen	0.681	
		Va.& Ra./s2	Va.& Ra./env2	wF1(o,t)	≈Va.& Ra.	OTE	Va.& Ra./d3	Bestgen	0.684	
		Va.& Ra./s1	Va.& Ra./env3	wF1(o,t)	Va.& Ra.	OTE	Va.& Ra./d1	Cai.&But.	0.657	
		Cai.&But.	Cai.&But.	wF1(o,t)	Cai.&But.	OTE	Va.&Ra./d4	Cai.&But.	0.401	

score reported by Vajjala and Rama, and seven reproduction scores from four different papers (Huber and Çöltekin 2020; Arhiliuc, Mitrović, and Granitzer 2020; Bestgen 2020; Caines and Buttery 2020). The conditions of measurement described above are given in columns 3–9; for space reasons the condition values are given as paper and method IDs; for example, *Test set = Va.&Ra./d2* means the original data was used, but the test set split was different. Finally, CV* scores are given in the last column.

Results show that system variant pairs that differ only in whether they use language information have very similar CV* scores. For example, mult-dep$^-$ (without language information) and mult-dep$^+$ (with language information) have a CV* of 4.5 and 4.39, respectively. This tendency holds for all such pairs (including the ones not shown here), indicating that using language information makes next to no difference to

reproducibility.[8] Results also show that the syntactic information is obtained/used in a way that is particularly reproducible, whereas the word 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 (mult-dom systems not included here) have the worst.

These insights can be used in different ways. The substantial differences in CV* seen here might prompt further code checking (and indeed there seem to have been issues with the mult-base code according to three of the reproduction papers). Knowing that the method for obtaining word embeddings is associated with particularly poor reproducibility might prompt an attempt to improve it. Automated CV* checking can be built into further development of the system to ensure good reproducibility of baseline variants.

4. Discussion

Is QRA all we need? QRA directly addresses the second of the questions (how similar are scores) mentioned at the start of Section 3 by computing degree of reproducibility as a function of scores. It also contributes to answering the first (how recreatable is the system) by providing pointers to which conditions of measurement (which aspect of system, training, compiling, evaluation, etc.) may be causing poor reproducibility. And it contributes to answering the third question (can the same conclusions be drawn) by providing a quantitative basis for deciding what is otherwise a subjective judgment. That is not to say that QRA does everything; in particular, conducting study-level comparisons (e.g., correlation between sets of scores, system rankings, and significant differences) provides additional insights (Fokkens et al. 2013).

What counts as a good level of reproducibility? This differs substantially between disciplines and contexts. In bio-science assays, precision (CV) ranges from <10 for enzyme assays, to 20–50 for in vivo and cell-based assays, and >300 for virus titre assays (AAH/USFWS n.d.). For NLP, typical CV ranges would have to be established over time, but it seems clear that we would expect them to be much lower (better) for metric-based measurements than for human-assessments. For wF1 measurements, the > 15 CV* scores above seem very high.

Does it matter how similar scores are? CV* provides a quantitative basis for deciding if the same conclusions can be drawn as discussed above, but looking at the degree of similarity of scores relative to similarity of conditions of measurement can also reveal important information, as exemplified by the systematic patterns in CV* found in the example application in the last section. Knowing what are expected degrees of reproducibility for a type of system or evaluation method helps pinpoint problems when those expectations aren't met.

If score differences in reproductions of the same system are larger than score differences reported as a new state of the art for the same task elsewhere, then that's a problem because it is unclear whether higher scores are in fact due to methodological improvements. If reproduction scores are, as would seem reasonable to assume, normally distributed, then overall, half of all reproduction scores should be better than

8 The high CV* for the baseline system may be due to an issue with the evaluation code (macro-F1 instead of weighted-F1), as reported by Bestgen (2020); Caines and Buttery (2020); and Huber and Çöltekin (2020).

the original scores, and half worse. But that is not the case: a recent survey (Belz et al. 2021a) showed that 60% of all reproductions that are different are in fact worse. In this way, examining differences between scores can facilitate higher-level insights, in this case, an apparent tendency to cherry-pick better results.

Differences in quality between studies. In the course of a reproduction study, it can happen that the reproducing authors discover what they consider errors in code or mismatches between the published description of the code and the code itself. In such cases, some judgment is involved: there is no point in reproducing an erroneous original study; if the code is different from the description, but not otherwise deemed problematic, a reproduction with corrected description might still make sense. Beyond actual errors, conditions of measurement capture the similarities and differences (some of which may be deemed a matter of quality) between different studies.

QRA in NLP workflows. In order to put metrological principles and quantified reproducibility assessment into practice, two things are needed at the field level: (i) recording conditions of measurement for all evaluations using a standard template; and (ii) routinely reporting CV* and conditions of measurement for reproduction studies. Ideally, the third element would be to conduct repeatability assessment as part of developing new evaluation methods, and usefully even as part of new applications of existing methods.

5. Conclusion

The reproducibility debate in NLP/ML has been framed in terms of exactly what information we need to share so that others are guaranteed to obtain the same metric scores, and initially the expectation was that "[r]eproducibility would be quite easy to achieve in machine learning simply by sharing the full code used for experiments" (Sonnenburg et al. 2007, page 2450). What is becoming clear, however, is that no matter how much of our code, data, and ancillary information we share, residual amounts of variation remain that are stubbornly resistant to being eliminated. A recent survey (Belz et al. 2021a) found that just 14% of the 513 original/reproduction score pairs analyzed 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. On the other hand, judging whether the same conclusions can be drawn on the basis of possibly different scores without a quantitative basis is prone to subjectivity and low agreement.

QRA offers an alternative by quantifying the closeness of results in a way that is comparable across different studies and can be used to establish, over time, expected levels of closeness for different types of systems and evaluation methods. Such expected CV* levels and the individual CV* results for given reproductions moreover provide a quantitative basis for deciding whether the same conclusions can be drawn, as well as providing information about the recreatability of systems and evaluation methods.

Reproducibility is one of the cornerstones of scientific research: Inability to reproduce results within accepted limits is, with few exceptions, seen as casting doubt on their validity. Building QRA into NLP workflows comes with an overhead, both in terms of recording information about systems and evaluations in a standardized form, and in terms of carrying out reproduciblity assessments. It can often seem in NLP that the performance-improvement paradigm and the pressure to maximize publications does not allow for any additional work not directly contributing to increasing performance or publications. In medicine and the physical sciences more generally it's inconceivable

that reproducibility assessment could be viewed as an optional extra. As NLP/ML matures as a science, perhaps it's time that we too insist on our results being verifiably reliable, including in the important sense of being reproducible.

References

AAH/USFWS. n.d. Assay validation methods: Definitions and terms. Aquatic Animal Health Program, U.S. Fish & Wildlife Service.

ACM. 2020. Artifact review and badging, Version 1.1, August 24, 2020. https://www.acm.org/publications/policies/artifact-review-and-badging-current

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

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

Baker, Monya. 2016. Reproducibility crisis. *Nature*, 533(26):353–366.

Belz, Anya. 2021. Quantifying reproducibility in NLP and ML. *arXiv preprint arXiv:2109.01211*.

Belz, Anya, 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.

Belz, Anya, Maja Popovic, and Simon Mille. 2022. Quantified reproducibility assessment of NLP results. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL'22)*, pages 16–28.

Belz, Anya, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021b. 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.

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

Branco, António, 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 REPROLANG 2020. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5539–5545.

Caines, Andrew 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.

Cohen, K. Bretonnel, Jingbo Xia, Pierre Zweigenbaum, Tiffany Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E. Hunter. 2018. Three dimensions of reproducibility in natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 156–165.

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

Fokkens, Antske, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1691–1701.

Huber, Eva 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.

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.

Mieskes, Margot, Karën Fort, Aurélie Névéol, Cyril Grouin, and Kevin Cohen. 2019. Community perspective on replicability in natural language processing. In *Proceedings of the International Conference on Recent Advances*

in Natural Language Processing (RANLP 2019), pages 768–775.

Pineau, Joelle. 2020. The machine learning reproducibility checklist v2.0. `https://www.cs.mcgill.ca/ jpineau/ReproducibilityChecklist.pdf`.

Rao, Calyampudi Radhakrishna. 1973. *Linear Statistical Inference and its Applications*. Wiley.

Rougier, Nicolas P., Konrad Hinsen, Frédéric Alexandre, Thomas Arildsen, Lorena A. Barba, Fabien C. Y. 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.

Schloss, Patrick D. 2018. Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *MBio*, 9(3):e00525-18.

Shimorina, Anastasia 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*.

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

Sonnenburg, Soren, Mikio L. Braun, Cheng Soon Ong, Samy Bengio, Leon Bottou, Geoffrey Holmes, Yann LeCunn, Klaus-Robert Muller, Fernando Pereira, Carl Edward Rasmussen, et al. 2007. The need for open source software in machine learning. *Journal of Machine Learning Research*, 8:2443–2466.

Vajjala, Sowmya 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.

Whitaker, Kirstie. 2017. The MT Reproducibility Checklist. Presented at the Open Science in Practice Summer School (`https://osip2017.epfl.ch/page-145979.html`) 2017.

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