Journal of Data Science 20 (3), 339–357
July 2022

DOI: 10.6339/22-JDS1054
Data Science in Action

Improving the Science of Annotation for Natural Language Processing: The Use of the Single-Case Study for Piloting Annotation Projects

Kylie Anglin[1,*], Arielle Boguslav[2], and Todd Hall[2]

[1]*Department of Educational Psychology, University of Connecticut, Storrs, CT, 06269, USA*
[2]*Department of Education Leadership Foundations and Policy, University of Virginia, Charlottesville, VA, 22902, USA*

Abstract

Researchers need guidance on how to obtain maximum efficiency and accuracy when annotating training data for text classification applications. Further, given wide variability in the kinds of annotations researchers need to obtain, they would benefit from the ability to conduct low-cost experiments during the design phase of annotation projects. To this end, our study proposes the single-case study design as a feasible and causally-valid experimental design for determining the best procedures for a given annotation task. The key strength of the design is its ability to generate causal evidence at the individual level, identifying the impact of competing annotation techniques and interfaces for the specific annotator(s) included in an annotation project. In this paper, we demonstrate the application of the single-case study in an applied experiment and argue that future researchers should incorporate the design into the pilot stage of annotation projects so that, over time, a causally-valid body of knowledge regarding the best annotation techniques is built.

Keywords *annotation; coding; single-case study; supervised machine learning; text classification*

1 Introduction

Text classification is playing an increasingly important role across many research domains. While traditional approaches to analyzing natural language data rely on trained personnel to read and annotate each document of interest, text classification methods only require hand-labeling text for a subset of the available documents. These data are then used to train an algorithm to automatically apply the annotation scheme to the remaining documents in the corpus. Notably, once the text classification algorithm has been trained, it can be applied again and again to additional documents at negligible cost. This feature makes text classification a powerful and efficient analytic tool when text is voluminous. However, the success of any text classification project depends on the amount and quality of training data available.

Given the importance of hand-labeled training data, computational linguists have begun to build a "science of annotation" advising researchers on the best methods of producing training data for natural language processing applications (Hovy and Lavid, 2010). Because of this work, researchers can find insightful and practical recommendations (Hovy and Lavid, 2010; Ide and

*Corresponding author. Email: kylie.anglin@uconn.edu.

Received December 11, 2021; Accepted June 20, 2022

Pustejovsky, 2017; Pustejovsky and Stubbs, 2012), as well as several studies empirically testing the impact of annotation techniques on annotator accuracy and efficiency. This includes research on the potential influence of annotator characteristics (Alyuz et al., 2021; Snow et al., 2008), of iterative consensus building among annotators (D'Mello, 2016), and of pre-annotation (Lingren et al., 2014). Yet there are many open questions, for example: if a classification task involves multiple labels, should annotators annotate texts for all labels at once or one at a time? How much context surrounding a given text should be provided to annotators? And, which interfaces best support annotators in annotating quickly and accurately? In many cases, answers to these kinds of questions will be highly dependent on annotator knowledge and experience, as well as the specific parameters of the task.

Ideally, researchers would make these decisions empirically during the pilot stage of annotation projects, before devoting significant resources to a sub-optimal annotation procedure. The gold standard for empirical decision making is the randomized control trial; researchers randomly assign annotators to one of two conditions and compare their resulting accuracy and efficiency. Unfortunately, smaller research teams rarely have a large enough pool of annotators for the randomized control trial to be feasible. Instead, empirical tests of annotation techniques commonly take one of two forms, both of which create challenges for identifying the causal effect of particular annotation methods. In one common approach, researchers may use a pre-post design where annotators use one method of annotation followed by an alternative method. Performance statistics are then compared across the time points. This design is straightforward but presents severe challenges for causal inference: namely, it is impossible to decipher whether changes in annotator performance are due to the new method of annotation, or due to increased annotator experience or any of a number of other time-varying confounders (Shadish et al., 2002). In another approach, researchers may split annotators into two groups and ask each group to use a different annotation method. Performance statistics are then compared across groups. However, if participants were not randomly assigned, or samples sizes are not large enough to ensure that groups are balanced on potentially confounding characteristics, then the causal impact of the annotation procedure cannot be differentiated from differences in performance due to annotator characteristics (Shadish et al., 2002).

We argue that the single-case study design addresses these causal inference challenges and offers a feasible solution for experimentation in annotation projects. The design controls for both time-varying and participant-varying confounders by switching the annotation procedure multiple times and comparing outcomes within (rather than across) participants (Kratochwill et al., 2013). If the annotation procedure is manipulated many times by the researcher, and the changes in participant performance track this pattern of manipulation, the researcher can conclude a causal relationship. The single-case study gets its name from the fact that the design can include as few as one participant. This makes it particularly well-suited for testing the efficacy of competing annotation techniques in fields like the social sciences where hand-labelling often requires domain-specific expertise. In addition to participating in an extensive training process, social science annotators are often required to have relevant professional and educational experiences that relate to the project's specific research area (Shaffer and Ruis, 2021). This limits the number of annotators that can be included in a given study. Yet, acquiring training data still requires substantial resource investment given the time that experts spend categorizing rich social constructs (Liu and Cohen, 2021). Social science researchers, then, need to know how to use annotator time effectively. In these cases, the single-case study offers a low-cost and causally valid solution for testing the comparative efficacy of multiple annotation procedures during the pilot stages of an annotation project. With the single-case study design, researchers can identify

the most effective techniques for their specific pool of annotators before devoting substantial resources.

In this article, we first provide an overview of the single-case study design for those who may not be familiar. Second, we review key decision points in annotation projects, highlighting points where the single-case study can aid in empirical decision making. Third, we illustrate the application of the design through an applied experiment testing two competing approaching to multi-label annotation projects. Finally, we discuss the generalizability of single-case study results and the strengths and weaknesses of the single-case study design for improving annotation science.

2 The Single-Case Study Design

Single-case study designs originated in psychology and date back to the field's founders (Perone and Hursh, 2013; Skinner, 1938; Watson, 1925). In contrast to the between-subject design, the single-case study relies on within-subject comparisons, where participants provide their own control data. The researcher assigns different treatment conditions to the same individual at different points in time while consistently measuring the outcome of interest. If the treatment assignment is manipulated many times by the researcher and the changes in outcomes track this pattern of treatment manipulation, the researcher concludes that the treatment caused the changes in outcomes. This conclusion is warranted when it is difficult to hypothesize confounders that would also produce the observed pattern of effects (Kratochwill et al., 2013). Conclusions from a single-case study are primarily drawn from visual analysis of graphs (Kratochwill et al., 2010). To provide evidence of a treatment effect, the graph should demonstrate an unlikely change in the pattern of data that correlates with the researcher's manipulation of the treatment condition (What Works Clearinghouse, 2019). A stylized example of a convincing single-case study is provided in Figure 1.

According to the What Works Clearinghouse (a governmental organization that rates the rigor of empirical evidence in education), the single-case study design is one of only three designs

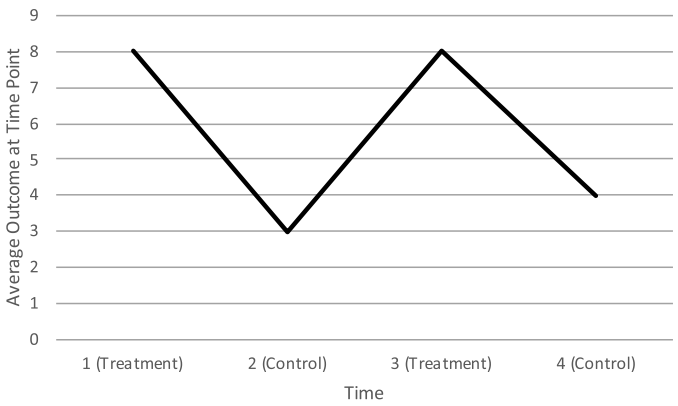


Figure 1: Stylized example of a single-case study design with one participant and a clear causal impact. Outcomes at each time point may either be single observations taken from the participant or average outcomes across many observations of the same participant. The strongest single-case studies are also replicated multiple times with more than one participant.

(including the randomized control trial and the regression discontinuity design) that meet high standards for causal evidence (2019). A strong single-case study has the following features: 1) the treatment is manipulated by the researcher, not by the study participants or the environment; 2) the outcome variables are measured systematically and consistently over time; and 3) there are at least three switches between each pair of treatment and control conditions studied (What Works Clearinghouse, 2019). Together, these features reduce the likelihood of confounding variables that produce the same pattern of effects as the manipulation in the treatment assignment.

The single-case study gets its name from the fact that the design can include as few as one participant. This feature makes it attractive for determining the impact of interventions when the participant pool is small. For example, the design is particularly popular in areas of psychology focused on evaluating treatments for rare or low-incidence diagnoses (Carbone VJ O'Brien et al., 2013). On the other hand, the limited number of participants also means that the single-case study can have potentially limited generalizability. While results can provide evidence of a causal effect for a single individual, this effect may or may not generalize beyond that individual. For this reason, researchers are expected to provide a comprehensive description of participants so that readers may consider the extent to which the impact of an intervention is likely to generalize to their population of interest (Kratochwill et al., 2013). Replicating the single-case study design with multiple participants can also provide stronger evidence of a generalizable effect. In other circumstances, a researcher may be most interested in identifying a causal effect for their own participants, without any need for generalization. This may occur in clinical cases when an individualized treatment must be chosen, or in annotation projects where the researcher wishes to choose the most efficient method of annotation for their specific set of annotators.

Given the single-case study's ability to generate causally valid individual treatment effects, the design can be very useful for researchers making decisions at the beginning of an annotation project. Annotation projects often include only a handful of annotators, making other causally valid designs, like the randomized control trial, infeasible. The single-case study offers a low-cost yet causally valid solution for these circumstances. On the other hand, there are a few disadvantages to using the single-case study to pilot annotation projects. The first is the limited generalizability of the design, discussed above. The second two limitations are more practical. First, while novice annotators are commonly expected to improve as they gain initial experience with the task (Donmez et al., 2010), changing the annotation procedures multiple times may slow this learning process. Second, depending on the outcome of interest, piloting an annotation project with the single-case study design may require some duplication of effort. For example, if researchers wish to measure inter-rater agreement as the outcome, multiple annotators will need to label the same documents throughout the course of the experiment. Without a pilot study, fewer documents may need to be annotated more than once, resulting in a greater number of labelled documents overall. Thus, piloting annotation with the single-case study is most appropriate for projects that require a large enough corpus of labelled documents that the initial start-up costs (in accuracy and resources) are worth the potential longer-term gains in annotator accuracy and speed resulting from choosing the best methods of annotation.

3 Key Questions in the Science of Annotation

Annotation is a complex and multi-part process. As a result, researchers are faced with many decisions in designing and implementing an annotation scheme. Here, we focus on two key decisions that can be answered empirically: Who should create the annotations? And how should

they do it? For the most part, we sidestep the question of *what* should be annotated, as that decision is wholly dependent on the research question at hand. We simply note there is broad agreement that 1) the annotated corpus needs to be representative of the population of interest (Manning and Schütze, 1999); and 2) that researchers should create a comprehensive *codebook* (also called a manual) that specifies the definitions of labels and provides examples (Hovy and Lavid, 2010). Because labels need to be theoretically valid and fit the data, creating the codebook is often an iterative process where the researcher moves back and forth between theory and data before finalizing the definitions (Auerbach and Silverstein, 2003; Hovy and Lavid, 2010). Helpfully, researchers can find substantial guidance on creating a codebook. See, for example, Hovy and Lavid (2010), Ide and Pustejovsky (2017), Chi (1997), and Shaffer and Ruis (2021).

3.1 Who Should Annotate?

One of the first decisions researchers need to make in an annotation project is who should create the annotations. Researchers may produce annotations themselves, identify content-area experts to produce the annotations, train undergraduate or graduate students (as is common in academic papers), or rely on untrained annotators from crowd-sourcing platforms like Amazon's Mechanical Turk (Geiger et al., 2020). There is a general understanding that the cost per annotation resulting from crowd-sourcing can be substantially less expensive than the cost per annotation resulting from content-area experts (Snow et al., 2008; Fort, 2016). However, it is also hypothesized that crowd-sourced annotations will be of lower quality. This hypothesis has been, at least partially, substantiated with empirical evidence. In a comparison of annotations created by expert annotators to those created by crowd-sourced workers, Snow et al. found higher agreement among expert annotators than between expert and non-expert annotators (2008). However, they also found that accuracy can be increased to the level of that achieved by experts by aggregating the annotations of multiple non-experts (2008). Importantly, the accuracy costs of relying on non-expert annotators will be very dependent on the specific annotation task. Snow et al., for example, conducted their tests on tasks requiring only general knowledge of the English language (2008). More specialized tasks may result in lower accuracy among non-experts. Where relevant, researchers may test this in their own data. Thankfully, there are few causal challenges in identifying the effect of one group of annotators versus another. This is because when testing the impact of different annotators, the researcher does not need to worry about annotator characteristics confounding the outcomes; differences between annotators are not confounders but instead the treatment of interest. Thus, so long as the researcher holds other features constant (like time and the annotation task), comparisons of outcomes across participant pools is valid.

3.2 How Should The Corpus Be Annotated?

After selecting the annotators, researchers need to determine how the annotators will produce their annotations. This involves selecting the annotation procedures and the annotation interface. Decisions will depend in large part on the type of classification task: binary, multi-label (where multiple categories may be applicable at a time), or multi-class (where there are multiple categories but only one may be applicable at a time). When it comes to multi-label classification tasks with longer documents, computational linguists are commonly advised to break a complex annotation project down into a series of simple micro-tasks, asking annotators to consider one label at a time and to view the text within just a small context window (Sabou et al., 2014; Hovy and Lavid, 2010). We call this approach to annotation the *simplified* annotation

procedure. Computational linguists commonly argue that simplifying an annotation scheme can increase efficiency by placing a lower cognitive load on annotators (Hovy and Lavid, 2010; Ide and Pustejovsky, 2017; Sabou et al., 2014). Though there is an accuracy cost to removing an utterance from its context (Samei et al., 2014), doing so also allows researchers to shorten the annotation task, which is hypothesized to increase annotator efficiency and accuracy enough to make up for performance lost due to lack of context. Hovy and Lavid argue, for example, that "though [the simplified annotation procedure] compromises on sentence context, [it] is both far quicker and far more reliable: annotators need to hold in mind just one set of alternatives, and become astonishingly rapid and accurate" (2010, p. 10). Further, by decomposing a task into short and simple yes or no questions, it becomes more feasible to obtain annotations from untrained annotators on crowd-sourcing platforms (Sabou et al., 2014). Breaking a multi-label task into multiple simple questions also has the added benefit of flexibility; when annotators annotate for all labels at once, the codebook becomes brittle. Changes to the annotation scheme would require re-annotating all utterances. The simplified approach allows for labels to be changed or edited without wasting substantial effort (White et al., 2019).

The simplified annotation scheme has been incorporated into many large-scale annotation projects. For example, this is the approach of the Decompositional Semantics Initiative, which decomposes complex linguistic concepts into "straightforward questions on binary properties that are easily answered" by untrained native speakers (White et al., 2016, p.1713). Similarly, the makers of the popular new annotation software, Prodigy, celebrate the software for allowing annotators to "focus on one task at a time" (Explosion AI, 2017). However, the simplified annotation approach is rarely taken by social scientists, either in traditional qualitative research or in text classification. In social science projects, annotators commonly consider one document at a time, annotating for every label in the codebook at once, an approach we call the *complex* annotation procedure (see, for example, D'Angelo et al., 2020; Loksa and Ko, 2016). The complex annotation procedure increases the cognitive load of annotation but also increases the information available to annotators. The applied experiment in this paper demonstrates how the single-case study design may be used to empirically assess the trade-offs between these two perspectives.

Finally, after specifying the annotation procedures, researchers need to identify the annotation interface, i.e., the software with which the annotators will interact. Ide and Pustejovsky identify many potential interfaces, including asking annotators to maintain a simple comma-separated-value file, contribute to a SQL database, or use a software specifically designed for annotation (2017). Neves and Ševa also provide an extensive review of annotation software based on technical criteria (including the cost and ease of installation), data criteria (including the input and output format of documents), and functionality criteria (including whether the software supports multi-label annotations and document-level annotations; 2021). Following these criteria, they recommend three programs that likely meet the needs of most users: WebAnno, brat, and FLAT. Unfortunately, however, there is currently little causally-valid evidence comparing the accuracy and efficiency of annotations resulting from competing interfaces. Helpfully, the single-case study design provides a key opportunity to affordably obtain such information.

4 Applied Experiment

In this section, we demonstrate how the single-case study design may be used to inform the development of annotation projects and to answer key questions in annotation science. Specifically,

we conduct two experiments. In the first experiment, we empirically assess two competing approaches to human annotation for multi-label annotation tasks with long documents: the *simple* and *complex* annotation procedures. In the complex approach, popular in the social sciences, annotators consider all labels at once and consecutively annotate text segments within the context of a full document. In the simplified approach, popular in computational linguistics, the annotation task is broken down into short and simple micro-tasks. Annotators view short text segments outside the context of the full document while annotating for one category at a time. Our first experiment is designed to determine which of these two approaches results in more efficient and accurate annotations. In a second follow-up experiment, we isolate the role of context and determine whether limiting the text annotators view increases or decreases efficiency and accuracy.

The study is situated within a broader educational research project focused on the efficacy of one-on-one coaching for improving teacher practice. The goal of the research is to use text classification to automatically monitor the strategies employed by coaches in their conversations with teachers and teachers-in-training. To this end, a coaching expert developed an annotation scheme and codebook by iteratively drawing on coaching research, practitioner resources, their professional experience receiving and providing coaching, and a random sample of coaching transcripts. The initial annotation scheme included over 30 potential strategies. For the purposes of text classification, we will initially focus on eight of the most common strategies: positive evaluation, observation, suggestion, instruction, demonstration, anticipation, practice, and encouragement. A description of these strategies, along with examples, is provided in the Supplementary Materials. In a single turn, a coach can employ as many as eight strategies or as few as zero. This means our project involves a multi-label classification task in which there are multiple categories (distinguishing it from a binary classification task) and many can apply at once (distinguishing it from a multi-class task).

4.1 Study Corpus and Participants

Our corpus of coaching conversations comes from prior studies of the impact of a short (5-minute) coaching intervention on teachers-in-training. For more details on the coaching intervention and its effects, see Cohen et al. (2020). All coaching conversations were recorded, professionally transcribed, and segmented by turns-of-talk. To pilot the annotation project, we randomly selected 30 coaching transcripts, 508 utterances in total. Then, we developed a gold-standard corpus; two coaching experts read the randomly selected transcripts and carefully labelled each coach utterance with the appropriate labels (agreement = 0.96, Krippendorff's alpha = 0.82). Because accuracy was the only priority in the creation of the gold standard corpus, the experts viewed each utterance within the context of the full transcript and took no steps to increase their own efficiency.

Four annotators were recruited through the university's centralized system for hiring undergraduate workers. The job was advertised to students across all schools and majors at the university. Applicants submitted a resume and short cover letter explaining their interest in the project and participated in a short video interview. While all four annotators had research experience and were in their third or fourth year of study, only two had prior teaching experience or a major within the school of education. Three out of four annotators had prior experience with annotation in qualitative research. This hiring and recruitment process followed the typical approach in social science research projects (Crittenden and Hill, 1971).

In the follow-up experiment designed to isolate the impact of context in the complex annota-

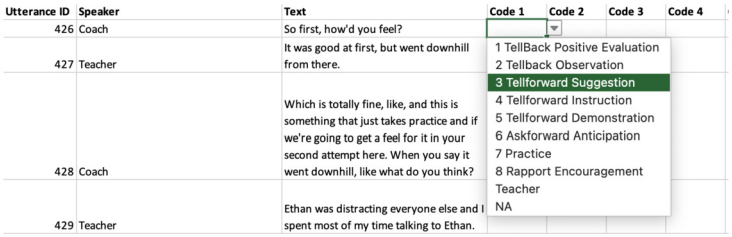


Figure 2: Complex annotation procedure interface. All coach and teacher utterances in a given transcript were included in the order in which they were spoken. Annotators considered one transcript at a time and all labels at once.

tion procedure, we sampled an additional 20 transcripts, 360 utterances in total. This experiment was conducted with three of the four annotators. (One annotator could not participate in the follow-up experiment.)

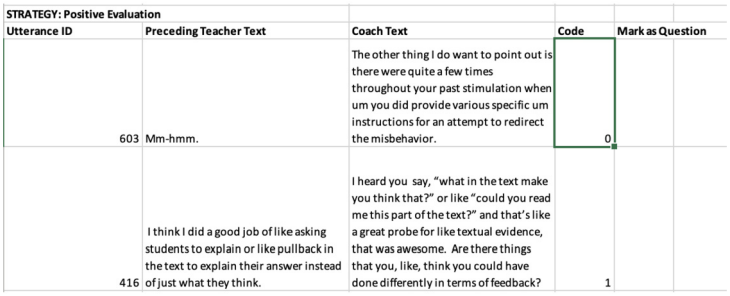
4.2 Annotation Procedures and Interface

In line with our annotators' prior technological experiences, we chose an interface implemented in Microsoft Excel because of its familiarity. Utterances were displayed in one column of the interface and the annotators entered their labels in a separate column (or columns). Specific annotation instructions depended on whether the annotators were annotating under the complex or simplified annotation procedure; we describe these details below.

Under the complex annotation procedure, annotators were asked to annotate one transcript at a time and to consider all coaching strategies at once. To this end, their annotation interface included one file per transcript. In each file, transcripts were formatted so that each row was a turn-of-talk. Turns-of-talk were kept in the order in which they were spoken, including both coach speech and teacher-in-training speech. For each coach utterance, annotators selected labels from a drop down menu containing the eight coaching strategies and an option for "None of the above." When appropriate, annotators could select subsequent labels in additional columns to right. When annotators finished annotating a transcript, they opened the next file to continue with the next transcript. For an example of this annotation interface, see Figure 2.

In the simplified annotation procedure, annotators were asked to consider one coaching strategy at a time. Thus, annotators were provided with one file per label (rather than one file per transcript). Again, each row was a turn-of-talk. However, turns-of-talk were presented in random order so that utterances were viewed with only the preceding teacher turn-of-talk as context. Annotators were then asked to enter a zero or one indicating whether the coach's speech was an exemplar of the target label. Once annotators finished annotating all utterances for one coaching strategy, they would open the next file and annotate the same utterances for the next coaching strategy. For an example of this annotation interface, see Figure 3.

For the purpose of the experiment, all turns-of-talk were annotated at least five times (for all eight labels). Each document was annotated by two hired annotators using the complex annotation scheme, two hired annotators using the simple annotation scheme, and at least one expert annotator providing the gold standard annotations.



STRATEGY: Positive Evaluation				
Utterance ID	Preceding Teacher Text	Coach Text	Code	Mark as Question
603	Mm-hmm.	The other thing I do want to point out is there were quite a few times throughout your past stimulation when um you did provide various specific um instructions for an attempt to redirect the misbehavior.	0	
416	I think I did a good job of like asking students to explain or like pullback in the text to explain their answer instead of just what they think.	I heard you say, "what in the text make you think that?" or like "could you read me this part of the text?" and that's like a great probe for like textual evidence, that was awesome. Are there things that you, like, think you could have done differently in terms of feedback?	1	

Figure 3: Simplified annotation procedure interface. Coach utterances were presented in randomized order along with the preceding teacher utterance. Annotators considered one label at a time.

4.3 Measures

For each annotation procedure, we developed analogous methods for measuring efficiency and validity. To assess annotator efficiency, annotators were asked to record their start and end time for each annotation file (either the time it took to annotate a transcript or the time it took to annotate all potential exemplars of a coaching strategy). We then converted these values into a measure of time spent per utterance, which served as our efficiency metric. In the complex annotation scheme, this was simply the average time it took an annotator to consider the appropriate labels for an utterance. In the simplified annotation scheme, this was the summation of the average time it took annotators to consider an utterance for all of the eight labels. Because the simplified scheme requires annotating the same utterance multiple times (here, eight times), a full picture of efficiency requires us to calculate total time spent annotating an utterance.

To assess validity, we measured the micro accuracy, precision, and recall of the resulting annotations under each procedure (calculating the metrics globally across all eight labels by counting the total number of true positives, false negatives, and false positives). Accuracy here is defined as annotator agreement with the gold-standard corpus. We measured accuracy by calculating the percent of correctly classified utterance-label pairs; because the annotators classified each turn-of-talk as representative – or not – of eight separate labels, it was possible for an annotator to accurately classify an utterance for one label, but incorrectly classify the utterance for a second label. Because our transcripts were imbalanced (no single strategy is present in more than 50% of the utterances, and some are present in less than 10%), it is also important to measure precision and recall. We measure precision by calculating the proportion of true positive labels out of all labels and measure recall by calculating the proportion of true positive labels that the annotator identified as such.

4.4 Study Design

We first randomly assigned four annotators to their starting condition (either the simplified or complex annotation procedure). After the first week of annotation, annotators were instructed to switch their method of annotation (from the simplified to the complex, or vice versa) at the beginning of each of the successive three weeks (see Table 1). We repeated this process in the follow-up experiment. In single-case study terms, this design is referred to as the ABAB design. It is the switching mechanisms that provide the study with high causal validity; if the impact of switching conditions is clear and conistent across each switch, then it is very

Table 1: Study design and annotation procedure assignments.

Annotator	Week 1	Week 2	Week 3	Week 4
1	Complex	Simplified	Complex	Simplified
2	Simplified	Complex	Simplified	Complex
3	Complex	Simplified	Complex	Simplified
4	Simplified	Complex	Simplified	Complex

difficult to hypothesize alternative explanations for the observed changes in outcomes. Thus, if the simplified annotation scheme increases (or decreases) annotation accuracy or efficiency, these changes can be causally attributed to the annotation condition. Our study design meets all of the What Works Clearinghouse standards for a causally-valid single-case study (What Works Clearinghouse, 2019).

4.5 Statistical Analysis

In each experiment, annotator efficiency is calculated each week for four weeks, resulting in four data points per annotator. Though this is a sufficient number of observations for causal inference using the typical single-case study graphs, it is an insufficient number of observations for statistical tests of significance. However, for precision, recall, and accuracy, we have over a thousand annotations for each annotator: enough to determine whether, for each participant, there is a statistically significant difference in these measures depending on the annotation condition. However, readers should be careful not to misinterpret these tests of significance. Statistical inference here is used to make inferences from a sample of utterances to a population of utterances, not from a sample of annotators to a population of annotators.

We use the following model:

$$\begin{aligned} Y_{ijk} = \beta_1 Simple_{ik} Annotator1_{ik} + \beta_2 Simple_{ik} Annotator2_{ik} + \\ \beta_3 Simple_{ik} Annotator3_{ik} + \beta_4 Simple_{ik} Annotator4_{ik} + \\ Week_i \beta_5 + Annotator_k \beta_6 + Label_j \beta_7 + \epsilon_{ijk}, \end{aligned} \tag{1}$$

where Y_{ijk} is a binary variable for whether a given turn-of-talk, i, was accurately annotated for label j, by annotator k; $Annotator_k$ is a vector of indicators for each of the four annotators; $Week_i$ a vector of indicators for each of the four weeks, and $Label_j$ a vector of indicators for each of the eight labels in the codebook. The coefficients of interest here are β_1 through β_4: the average impact of the simplified annotation procedure for each of the four annotators.

We also summarize the results across our four participants using the following model:

$$Y_{ijk} = \beta_1 Simple_{ik} + \beta_2 Week_i + Annotator_k \beta_3 + Label_j \beta_4 + \epsilon_{ijk}, \tag{2}$$

where Y_{ijk} is a binary variable for whether a given turn-of-talk, i, was accurately annotated for label j by annotator k; $Annotator_k$ is a vector of indicators for each of the four annotators; $Week_i$ a continuous variable for the week; and $Label_k$ a vector of indicators each of the eight labels. The coefficient of interest here is β_1, the average impact of the simplified annotation scheme across all four annotators and eight labels.

All models were estimated using the `statsmodels` (Seabold and Perktold, 2010) and `pandas` (McKinney, 2010) packages in `Python 3.10.0` (Van Rossum and Drake, 2009). Figures were produced using `matplotlib` (Hunter, 2007) and `seaborn` (Waskom, 2021).

4.6 Results

The effect of the simplified annotation procedure on annotator efficiency is presented in Figure 4. The figure demonstrates that the simplified annotation procedure took roughly twice as long as the complex annotation procedure to produce the same number of labels. In the complex procedure, annotating an utterance for all of the eight labels at once took 35.5 seconds on average. In the simple annotation procedure, annotating an utterance for a single label took only 8.5 seconds, but this approach requires annotators to read each utterance eight separate times, thus, requiring 68 seconds per utterance to produce the same number of labels as the complex procedure (the sum of the average time spent on each of the eight individual labels). In other words, although reviewing an utterance for a single label took annotators less time than reviewing the utterance for multiple labels, the time spent was not reduced by a factor of eight, which would be required to make the simplified annotation more efficient than the complex procedure in this case.

From a single-case study point of view, Figure 4 provides convincing evidence of causality; manipulation of the treatment condition here is associated with a consistent change in the dependent variable. The effect is visually obvious at each switch in the treatment conditions and is replicated for every participant in the study. Each individual takes more time when annotating under the simplified annotation procedure than when annotating under the complex annotation procedure. For one individual, this effect is small (Annotator 2), while for the others it is much larger. Crucially, it is very difficult to provide any alternative explanation for the change in times given that the effect is demonstrated at every switch in treatment condition and for every annotator. No other confounding variable is likely to display this same pattern of effects.

Unlike Figure 4, Figure 5 does not demonstrate a strong or consistent impact of the simplified annotation procedure on accuracy. While the simplified annotation procedure causes a decrease in accuracy for one annotator (Annotator 2), the effect is not convincingly replicated with the other annotators. In Table 2, we summarize the average accuracy for each annotator under the two annotation schemes using Equation 1. While annotator accuracy was high

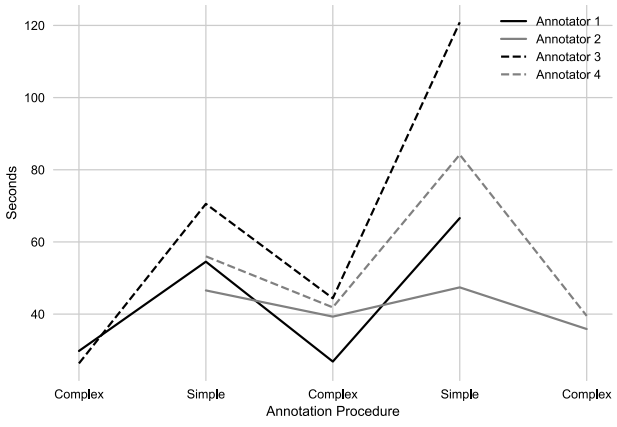


Figure 4: Average total annotation time per utterance as a function of the annotation procedure. Under the complex annotation scheme, this is the average time it took the annotators to consider the relevance of the eight labels all at once for a given utterance. Under the simplified annotation scheme, this is the average total time it took annotators to consider the relevance of the eight individual labels, one at a time.

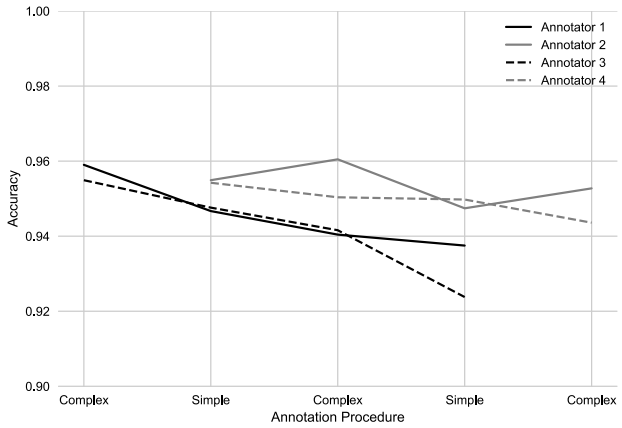


Figure 5: Accuracy as a function of the simplified versus complex annotation procedure. The y-axis of the figure displays average accuracy across all utterance-label pairs for each annotator and each week.

Table 2: Impact of the simplified annotation procedure on accuracy, precision, and recall.

	Accuracy ($M = 0.95$)	Precision ($M = 0.83$)	Recall ($M = 0.75$)
Annotator 1*Simplified Procedure	−0.003	−0.003	0.041
	(0.008)	(0.037)	(0.041)
Annotator 2*Simplified Procedure	−0.011	−0.035	−0.119**
	(0.008)	(0.036)	(0.036)
Annotator 3*Simplified Procedure	−0.006	−0.052	0.103**
	(0.009)	(0.04)	(0.04)
Annotator 4*Simplified Procedure	−0.001	−0.043	−0.014
	(0.008)	(0.038)	(0.04)
Average Impact Across Annotators	−0.005*	−0.035**	0.003
	(0.003)	(0.012)	(0.015)

Note. $N = 508$. The first four rows of the table represent the impact of the simplified annotation procedure on each of the four annotators' accuracy, precision, and recall, estimated using Equation 1. The final row represents the average impact of the simplified annotation procedure across all four annotators, estimated using Equation 2. The average annotator accuracy, precision, and recall, regardless of annotation condition, M, is presented in parentheses below the column titles. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

across the board (95% on average), there are no statistically significant differences in accuracy by annotation procedure for any of the four annotators. When we aggregate these results across annotators using Equation 2, the overall impact of the simplified annotation scheme is a small, but significant, decrease in accuracy by half a percentage point.

Given the imbalanced nature of our data set, a full understanding of the impact of the simplified annotation procedure requires an analysis of precision and recall. Figure 6 demonstrates that the simplified annotation procedure caused a decrease in precision for three out of four annotators: on average, a statistically significant negative effect of 3.5 percentage points across all four annotators. On the other hand, Figure 7 demonstrates a heterogeneous impact of

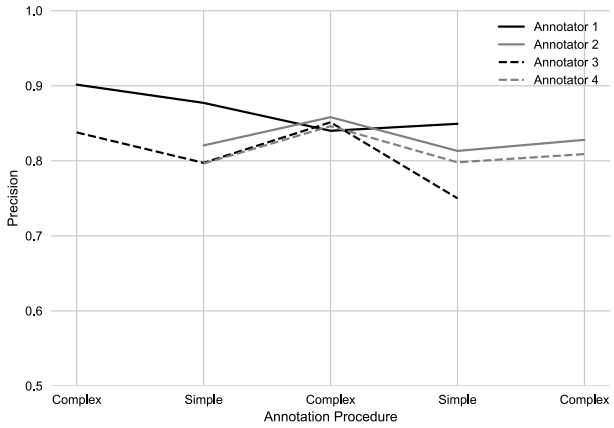


Figure 6: Precision as a function of the simplified versus complex annotation procedure. The y-axis of the figure displays average precision across all utterance-label pairs for each annotator and each week.

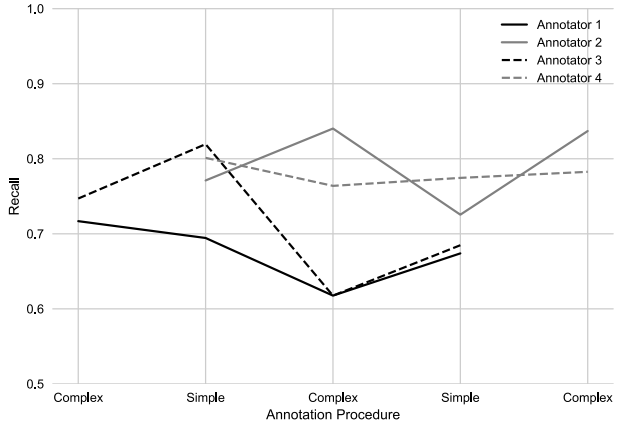


Figure 7: Recall as a function of the simplified versus complex annotation procedure. The y-axis of the figure displays average recall across all utterance-label pairs across all utterance-label pairs for each annotator and each week.

the simplified annotation scheme on recall. While two annotators experienced substantial and consistent impacts of the simplified annotation procedure, these effects are in opposite directions (-12 percentage points for Annotator 2 and $+10$ percentage points for Annotator 3; see Table 2). These two effects counterbalance one-another, resulting in a very small, non-significant, effect for recall overall. Taken together, the simplified annotation procedure increases the time spent annotating and reduces precision, with only a negligible negative impact on overall accuracy.

4.7 Isolating the Role of Context

Compared to the complex annotation procedure, the simplified annotation procedure is different in two keys ways: 1) it asks annotators to review an utterance for a single label at a time; and 2) it provides annotators with less context surrounding an utterance. The previous results demonstrated that the simplified annotation procedure was less efficient and resulted in

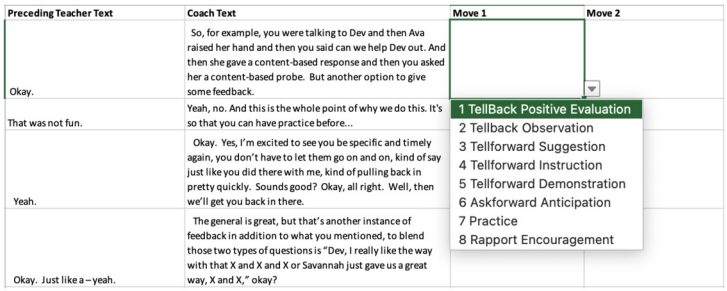


Figure 8: Out-of-context annotation interface. Coach utterances were presented in randomized order along with the preceding teacher utterance. Annotators considered all labels at once.

annotations with a lower rate of precision. To isolate the role of context in the these results, we conduct a short follow-up experiment which only varies the context provided to annotators. In one condition, annotators again annotate using the complex procedure ("complex"). In the other condition, annotators still consider all labels at once, but view the utterances in a random order with only the preceding utterance for context ("out-of-context"). For an example of the out-of-context interface, see Figure 8. As in the previous design, annotators switched conditions each week.

Figure 9 demonstrates that there is no substantial or consistent efficiency difference for either condition. Thus, context was mostly irrelevant in determining the amount of time annotators took to produce annotations. There is also no consistent impact for accuracy or precision. However, the reduced context caused annotators to produce annotations with lower recall (by four percentage points; see Figure 10 and Table 3). Taken together with the results of the prior experiment, these results suggest that the increase in efficiency of the complex annotation procedure in the first experiment was due to annotators considering all labels at once, not the amount of context provided. Differences in precision in the complex and simplified procedures are also likely due to differences in labelling rather than context (given the null impact of context on accuracy and precision in the follow-up experiment). The impact of context versus labelling procedures on recall, however, is more complex. While decreasing context decreases recall (as demonstrated in the follow-up experiment), the results of these experiments suggest that labelling one category at a time increases recall, cancelling out the negative effects of reduced context on recall in the simplified annotation procedure.

5 Interpreting the Results of the Applied Experiments

The above applied experiments tested two key questions in the design of a multi-label annotation task with long documents: should annotators annotate one label at a time, or all at once? And, what amount of context should annotators use to interpret each text segment? Given the results above, we determined that the best procedure for the annotators in this study is to annotate for all labels at once within the context of a full document (i.e., the complex annotation procedure). We did not find any efficiency benefits resulting from the simplified annotation procedure. In total, it took annotators twice as long to annotate the same data using the simplified annotation procedure than using the complex annotation procedure. Whatever cognitive speed was gained by requiring annotators to only consider one label at a time was not enough to outweigh the time

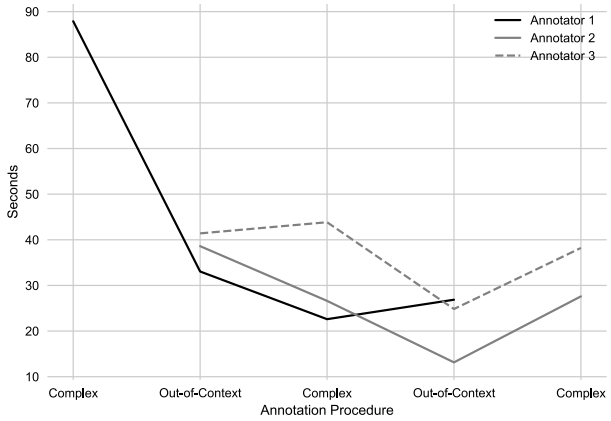


Figure 9: Average total annotation time per utterance as a function of the context provided to annotators. In both procedures, annotators considered eight labels at once, but while the "Complex" procedure displayed all utterances in order, the "Out-of-Context" procedure displayed coach utterances in randomized order. The y-axis of the figure displays the total average annotation time per utterance.

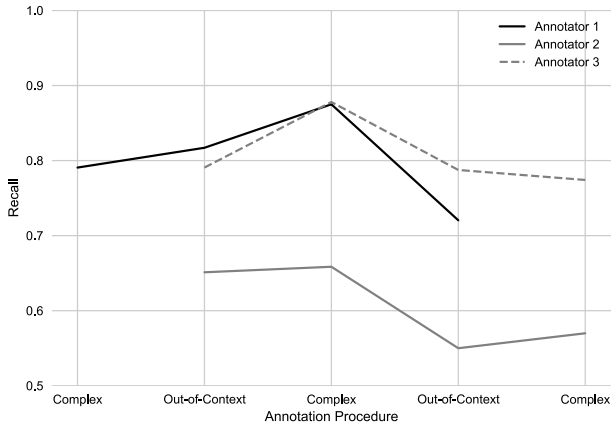


Figure 10: Recall as a function of the context provided to annotators. In both procedures, annotators considered eight labels at once, but while the "Complex" procedure displayed all utterances in order, the "Out-of-Context" procedure displayed coach utterances in randomized order. The y-axis of the figure displays average accuracy across all utterance-label pairs.

it took to consider the same utterance multiple times. When considering precision and recall, our results again suggest that, in this case, the complex procedure is preferable. The simplified annotation procedure reduced precision in the main experiment and the loss of context reduced recall in the follow-up experiment.

Of course, readers should be careful in their consideration of whether the findings of this study generalize to their own context. In particular, there are two dimensions along which generalizability should be considered. First, our study was conducted with undergraduate research assistants, three of whom had prior experience with annotation in other qualitative studies across the university. We might hypothesize that reducing cognitive load is more important when anno-

Table 3: Impact of the lack of context when annotating for multiple labels at once on annotator accuracy, precision, and recall.

	Accuracy ($M = 0.95$)	Precision ($M = 0.85$)	Recall ($M = 0.74$)
Annotator 1*Out-of-Context	−0.01	0.033	−0.045
	(0.008)	(0.041)	(0.045)
Annotator 2*Out-of-Context	0.005	−0.029	−0.03
	(0.009)	(0.044)	(0.052)
Annotator 3*Out-of-Context	0.005	−0.009	−0.054
	(0.009)	(0.048)	(0.047)
Average Impact Across Annotators	−0.002	0.007	−0.043*
	(0.003)	(0.016)	(0.022)

Note. $N = 360$. The first three rows of the table represent the impact of the lack of context, when annotating for eight labels at once, on accuracy, precision, and recall for each of three annotators, estimated using Equation 1. The final row represents the average impact of lack of context across all three annotators, estimated using Equation 2. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

tators lack experience or knowledge of the study context. While our annotators were not content area experts, they were also not novices to the same degree as annotators hired through crowd-sourcing platforms like Amazon Mechanical Turk. However, we believe this project has optimistic implications for the use of crowd-sourcing for social science text classification projects. Crowd-sourcing platforms *necessitate* short simplified annotation tasks. MTurkers, for example, expect each task to take a matter of seconds (Sabou et al., 2014). Previous research has demonstrated that crowd-sourced annotators can compete with the accuracy of more traditional annotators (Snow et al., 2008), however, this research does not address the potential loss of accuracy that comes with altering the annotation task so that it may be crowd-sourced. This study demonstrates that while simplifying an annotation task and taking excerpts outside of their larger context may reduce accuracy slightly, it is not to such a degree that social science researchers need to dismiss crowd-sourcing as a possibility.

Second, the relative trade-offs of the simplified and complex annotation procedures are likely to depend on the annotation scheme itself. In particular, the amount of text context required for sufficient accuracy will depend on the labels in the codebook. Labels which depend on information provided earlier in conversation will necessitate large context windows. Further, the benefits of the simplified annotation scheme are likely to vary by the number of labels in the codebook; as the number of labels increase, so to does the number of times an annotator must re-read an utterance. On the other hand, we can imagine there is some number of labels for which it becomes impossible for an annotator to remember all the definitions. At this point, asking annotators to consider one label, or groups of a labels, at a time may be necessary. We suggest that in cases of uncertainty, researchers should conduct their own tests. A key strength of the single-case study design is that such tests can be completed quickly and at relatively low cost.

6 Conclusion

This study demonstrates a straightforward and low-cost approach to testing methods of maximizing annotator efficiency and accuracy: the single-case study design. Given the limited number of participants required to make causal inferences, the single-case study design is well-suited to testing competing procedures or interfaces when annotation projects have only a few annotators. While the randomized control trial would be preferable in the case where an annotation project includes many annotators (say, close to 30), in our experience, researchers rarely hire that many annotators outside the context of crowd-sourcing. The single-case study design, on the other hand, can be valid with as few as one annotator. Thus, researchers can pilot annotation procedures quickly and cheaply, while also obtaining findings with high causal validity. Though each single-case study may only generalize to a subset of annotation projects, the relatively low cost of the design means that replicating findings across various contexts is feasible. Thus, we encourage researchers to use the single-case study both to inform their own annotation projects and to iteratively improve the evidence base regarding best practices in human annotation.

In the past, many text classification papers have neglected to give human annotations the consideration they are due (Geiger et al., 2020). Despite growing calls for researchers to document the origins and appropriate uses of training data in data statements or data sheets (Gebru et al., 2021; Bender and Friedman, 2018), many papers today still fail to report key information on how their training data were obtained (Geiger et al., 2020). Because of this, some researchers have deemed human-annotated corpora, the "hidden pillars" of natural language processing (Fort, 2016, p. 9). In this paper, we argue that researchers should respond to calls for increased attention to annotation quality by incorporating causal evidence into decision-making when designing annotation projects. If human annotations are the "hidden pillars" of text classification, we believe that we can increase the strength and visibility of these pillars through an increased focused on empirical, causally-valid decision-making in annotation.

Supplementary Material

The Supplementary Material includes all of the scripts and data files necessary to reproduce the results of this paper. We also include the codebook used by our annotators.

References

Alyuz N, Aslan S, SK Nachman L D, Esme AA (2021). Annotating Student Engagement Across Grades 1–12: Associations with Demographics and Expressivity. In: *Lecture Notes in Computer Science*, 42–51. Springer International Publishing.

Auerbach C, Silverstein LB (2003). *Qualitative Data: An Introduction to Coding and Analysis*, volume 21. NYU press.

Bender EM, Friedman B (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. In: *Transactions of the Association for Computational Linguistics*, volume 6, 587–604. MIT Press.

Carbone VJ O'Brien L, Sweeney-Kerwin EJ, Albert KM (2013). Teaching Eye Contact to Children with Autism: A Conceptual Analysis and Single Case Study. *Education and Treatment of Children*, 36(2): 139–159.

Chi MT (1997). Quantifying Qualitative Analyses of Verbal Data: A Practical Guide. *Journal of the Learning Sciences*, 6(3): 271–315.

Cohen J, Wong V, Krishnamachari A, Berlin R (2020). Teacher Coaching in a Simulated Environment. *Educational Evaluation and Policy Analysis*, 42(2): 208–231.

Crittenden KS, Hill RJ (1971). Coding Reliability and Validity of Interview Data. *American Sociological Review*, 36(6): 1073.

D'Angelo ALD, Ruis AR, Collier W, Shaffer DW, Pugh CM (2020). Evaluating How Residents Talk and What it Means for Surgical Performance in the Simulation Lab. *The American Journal of Surgery*, 220(1): 37–43.

D'Mello S (2016). On the Influence of an Iterative Affect Annotation Approach on Inter-Observer and Self-Observer Reliability. *IEEE Transactions on Affective Computing*, 7(2): 136–149.

Donmez P, Carbonell J, Schneider J (2010). A Probabilistic Framework to Learn from Multiple Annotators with Time-Varying Accuracy. *Society for Industrial and Applied Mathematics.*

Explosion AI (2017). Prodigy: A New Tool for Radically Efficient Machine Teaching. https://explosion.ai/blog/prodigy-annotation-tool-active-learning.

Fort K (2016). *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. John Wiley & Sons.

Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Iii HD, et al. (2021). Datasheets for Datasets. In: *Communications of the ACM*, volume 64, number 12, 86–92. Association for Computing Machinery (ACM).

Geiger RS, Yu K, Yang Y, Dai M, Qiu J, Tang R, et al. (2020). *Garbage in, Garbage out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?*. ACM.

Hovy E, Lavid J (2010). Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *International Journal of Translation*, 22(1): 25.

Hunter JD (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3): 90–95.

Ide N, Pustejovsky J (2017). *Handbook of Linguistic Annotation*, volume 1. Springer.

Kratochwill TR, Hitchcock J, Horner RH, Levin JR, Odom SL, Rindskopf DM, et al. (2010). *Single-Case Designs Technical Documentation. Technical report, What Works Clearinghouse.*

Kratochwill TR, Hitchcock JH, Horner RH, Levin JR, Odom SL, Rindskopf DM, et al. (2013). Single-Case Intervention Research Design Standards. *Remedial and Special Education*, 34(1): 26–38.

Lingren T, Deleger L, Molnar K, Zhai H, Meinzen-Derr J, Kaiser M, et al. (2014). Evaluating the Impact of Pre-Annotation on Annotation Speed and Potential Bias: Natural Language Processing Gold Standard Development for Clinical Named Entity Recognition in Clinical Trial Announcements. *Journal of the American Medical Informatics Association*, 21(3): 406–413.

Liu J, Cohen J (2021). Measuring Teaching Practices at Scale: A Novel Application of Text-as-Data Methods. *Educational Evaluation and Policy Analysis*, 016237372110092.

Loksa D, Ko AJ (2016). The Role of Self-Regulation in Programming Problem Solving Process and Success. In: *Proceedings of the 2016 ACM Conference on International Computing Education Research*. ACM.

Manning CD, Schütze H (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge.

Neves M, Ševa J (2021). An Extensive Review of Tools for Manual Annotation of Documents. *Briefings in Bioinformatics*, 22(1): 146–163.

Perone M, Hursh DE (2013). *Single-Case Experimental Designs*. ISBN: 143381112X. American Psychological Association.

Pustejovsky J, Stubbs A (2012). *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*. O'Reilly Media, Inc.

Sabou M, Bontcheva K, Derczynski L, Scharl A (2014). Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA)*, 859–866. 2014.

Samei B, Olney AM, Kelly S, Nystrand M, D'Mello S, Blanchard N, et al. (2014). Domain Independent Assessment of Dialogic Properties of Classroom Discourse. In: *Proceedings of the 7th International Conference on Educational Data Mining*, 4.

Seabold S, Perktold J (2010). Statsmodels: Econometric and Statistical Modeling with Python. In: *9th Python in Science Conference*.

Shadish WR, Cook TD, Campbell DT (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

Shaffer DW, Ruis AR (2021). How We Code. In: *ICQE 2020* (S Lee, AR Ruis, eds.). Springer, Malibu, CA.

Skinner BF (1938). *The Behavior of Organisms: An Experimental Analysis. BF Skinner Foundation*.

Snow R, O'Connor B, Jurafsky D, Ng A (2008). Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254–263. Association for Computational Linguistics, Honolulu, Hawaii.

Van Rossum G, Drake FL (2009). *Python 3 Reference Manual*. CreateSpace, Scotts, Valley, CA.

Waskom ML (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, 6(60): 3021.

Watson JB (1925). Experimental Studies on the Growth of the Emotions. *The Pedagogical Seminary and Journal of Genetic Psychology*, 32(2): 328–348.

McKinney W (2010). Data Structures for Statistical Computing in Python. In: *Proceedings of the 9th Python in Science Conference* (S van der Walt, J Millman, eds.), 56–61.

What Works Clearinghouse (2019). What Works Clearinghouse Standards Handbook: Version 4. *U.S. Department of Education's Institute of Education Sciences (IES)*, 1–17.

White AS, Reisinger D, Sakaguchi K, Vieira T, Zhang S, Rudinger R, et al. (2016). Universal Decompositional Semantics on Universal Dependencies. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1713–1723.

White AS, Stengel-Eskin E, Vashishtha S, Govindarajan V, Reisinger DA, Vieira T, et al. (2019). The Universal Decompositional Semantics Dataset and Decomp Toolkit. arXiv preprint: https://arxiv.org/abs/1909.13851.