Automated Scoring of Tablet-Administered Expressive Language Tests

Robert Gale, Julie Bird, Yiyi Wang, Jan van Santen, Emily Prud’hommeaux, Jill Dolata, and Meysam Asgari

1 Center for Spoken Language Understanding, Oregon Health & Science University (OHSU), Portland, Oregon, USA
2 Boston College, Chestnut Hill MA, USA
† These authors have contributed equally to this work and share senior authorship.

Correspondence*: Robert Gale
OHSU, Mail Code: CSLU
840 SW Gaines Street
Portland, OR 97239
galer@ohsu.edu

ABSTRACT

Speech and language impairments are common pediatric conditions, with as many as 10% of children experiencing one or both at some point during development. Expressive language disorders in particular often go undiagnosed, underscoring the immediate need for assessments of expressive language that can be administered and scored reliably and objectively. In this paper, we present: (1) our tablet-based child language assessment instrument; (2) a dataset we have collected with this instrument; and (3) a set of highly accurate computational models we have developed for scoring responses to several expressive language tasks administered with this instrument. In our assessment framework, instructions and stimuli are presented to the child on a tablet computer, which records the child’s responses in real time, while a clinician controls the pace and presentation of the tasks using a second tablet. The recorded responses for four distinct expressive language tasks (expressive vocabulary, word structure, recalling sentences, and formulated sentences) are then scored using traditional paper-and-pencil scoring and using machine learning methods relying on a deep neural network-based language representation model. All four tasks can be scored automatically from both clean and verbatim speech transcripts with very high accuracy at the item level (83 – 99%). In addition, these automated scores correlate strongly and significantly (r = 0.78 – 0.99, p < 0.001) with manual item-level, raw, and scaled scores. These results point to the utility and potential of automated computationally-driven methods of both administering and scoring expressive language tasks for pediatric developmental language evaluation.

Keywords: speech, expressive language, language disorders, assessment, automated scoring, neural language models
1 INTRODUCTION

Untreated and undiagnosed developmental language disorder (DLD) is prevalent in young children (Tomblin et al., 1997; Conti-Ramsden et al., 2006; Grimm and Schulz, 2014; Rosenbaum and Simon, 2016) and can have serious behavioral and educational consequences (Clegg et al., 2005). Wide-reaching language assessment is urgently needed not only for early identification of DLD but also for planning interventions and tracking the efficacy of these interventions. Such efforts, however, add strain to scarce and overtaxed clinical resources. To address this challenge, clinicians, educators, and researchers have begun to explore alternatives to standard assessment paradigms that can be more easily and more reliably administered and scored.

Conventional language test administration and scoring is a labor-intensive and time-consuming task relying on significant clinical expertise. Assessment is typically conducted during a clinical visit by a speech language pathologist (SLP) using a battery of standardized language tests in addition to criterion-referenced and dynamic measures. Responses are scored using paper score sheets, which are marked in real time and then later reviewed by the clinician. In addition to the practical overhead required to administer such assessments, the scoring of these tests can suffer from intra- and inter-rater variability (Denman et al., 2017). Although some progress has been made in developing automatic assessments for receptive language (i.e., language comprehension), there is no automatic assessment that elicits and analyzes spoken responses to expressive language tasks Marble-Flint et al. (2020, 2019). A sizeable fraction of children with DLD have primary difficulty with expressive communication (Tomblin et al., 1997). The nature of an individual’s language disorder impacts etiology, intervention, and psychiatric sequelae (Boyle et al., 2010; Yew and O’Kearney, 2013). Given that computerized administration ought to mimic conventional administration, spoken responses must be used in any proposed computerized assessment, just as they are in many subtests that contribute to general language composite scores (e.g., Semel et al. (2003)). While language assessment for children has predominantly utilized conventional face-to-face administration, automated testing could increase clinician efficiency, access to services, standardization of administration and scoring, and even child interest (García Laborda, 2007; Noland, 2017).

In this paper, we investigate the utility of a computerized tablet-based child language assessment instrument, modeled after the Clinical Evaluation of Language Fundamentals (CELF-4) (Semel et al., 2003). Using child language data collected both with this computerized instrument and with standard paper-and-pencil administration, we demonstrate the accuracy and feasibility of an automated scoring system for four expressive language tasks. The scoring methodology fine-tunes representation learning models to predict the score for responses to individual test stimuli, from which raw and scaled scores can be derived. In contrast to models that take manually-crafted measures of language as input, our deep neural network (DNN)-based model directly estimates scores directly from the transcripts. We find that our computerized scoring system yields very high item-level accuracy and summary score correlations when applied to both clean and verbatim speech transcripts. These results demonstrate the promise of computerized approaches to scoring expressive language tasks, which in turn can support clinicians tasked with diagnosis and extend the reach of services for children with developmental language disorder.

2 MATERIALS AND METHODS

2.1 Data

2.1.1 Participants

Participants in this analysis include 107 English-speaking children aged 5 to 9 years. These children represent the subset of participants from our larger study who completed automatic test administrations.
Our participants and their families provided informed consent, according to our institutional review board policies. The children’s demographic information is presented in Table 1. In order to ensure the utility of our system for a diversity of child ability, participants included children with Autism Spectrum Disorder (ASD, n=20), Attention-Deficit Hyperactivity Disorder (ADHD, n=19), Developmental Language Disorder (DLD, n=22), and typical development (TD, n=46). Diagnoses were confirmed via a combination of parent report of medical diagnosis, parent report of special education eligibility criteria, and expert review of parent-provided developmental history.

2.1.2 Data Collection

To match the individual subtest goals for assessment of expressive language, we first constructed candidate items using individual CELF-4 stimuli as template stimuli. Each of our four automated language measures directly corresponds to one of the subtests of the CELF-4, and items have been matched to these templates in terms of word class, vocabulary statistics (age of acquisition and word frequency), sentence structure, and Flesch-Kincaid readability scores. If an automatic text generation procedure had been used to create items solely under these constraints, the resulting items would not be guaranteed to be meaningful, particularly to children. Therefore, our items were created via automatic text generation and with the supervision of a licensed speech-language pathologist, who used the automatic language metrics to create a list of possible options and selected items based on appropriate context. For example, many prompts in the original CELF-4 were content-specific to school, classrooms, and peer and family relationships. We verified our selections by informal inspection of meaningfulness by all team members, including clinical experts in pediatric speech-language pathology, psychology, and psychiatry.

The automated administration was conducted using a custom-built iPad application. The application presented the stimuli for four subtests, mimicking the original presentation prompts as closely as possible, and recorded the audio of children’s spoken responses. The pacing and presentation of tasks and stimuli were controlled by a clinician using a second iPad. Trained examiners transcribed the responses and scored them according to conventional pen-and-paper rules.

2.1.3 Expressive Language Tasks

Four subtests are included in this study: expressive vocabulary (EV), word structure (WS), recalling sentences (RS), and formulated sentences (FS). In the EV subtest, the child views an image and must verbally name the person, object, or activity depicted in that image. Responses are scored with full credit (2) if correct, partial credit (1) if the response is not incorrect but not specific enough (e.g., “fruit” for “lime”), and no credit (0) if the response is entirely incorrect. The test is discontinued if there are seven consecutive scores of zero.

WS is used to assess a child’s grasp of inflectional morphology. A child is generally asked to complete a sentence after being given a prompt and shown a picture. Targets include morphemes such as verb tense inflections, possessives, plurals, and comparatives. An example item is, “the woman is fixing the car; here is the car that the woman... [fixed].” Scores for this subtest are simply correct (1) or incorrect (0), and there is no discontinuation rule.

In the FS subtest, the child hears a target word and views a photograph of a scene, and must produce a complete sentence about the scene using the target word. The targets include a variety of word classes and increase in their syntactic requirements, including more challenging targets such as, “safely” and “because.” Responses are scored as 0,1, or 2, depending on the use of target word, grammatical correctness, and meaningful content. The test is discontinued if there are five consecutive scores of zero.
In the RS subtest, the child hears a sentence once (along with the video of a person saying the sentence in our automated task) and is asked to repeat it verbatim. Responses are scored on a scale of 0-3 based on the number of errors (omissions, substitutions, transpositions). Repetitions are not counted as errors. The test administration is discontinued after five consecutive scores of zero.

2.1.4 Data Scoring and Processing

During automated testing, examiners wrote down the child’s response to the prompt. The audio recordings from the iPad application were later transcribed by research assistants, resulting in two transcripts for each audio response. We note that the content of these two transcripts often diverges. The response transcribed by the examiner in real time typically consists only of the word, phrase, or sentence to be scored, excluding any other commentary the child might provide. We refer to these transcripts as clean transcripts. The transcripts generated by the research assistants included all speech produced by the child that was recorded by the iPad, including comments and prefatory content. We refer to these transcripts as verbatim transcripts.

When presented in the expressive vocabulary task with a picture of a pirate, a child might respond “I think that one’s a pirate.” The verbatim transcript would include the entire utterance, while the clean transcript would include only the single target word “pirate.” For the verbatim transcripts, we removed a limited number of known filler words like “hmm,” “ow,” and “uh-uh,” as well as special symbols used during transcription for non-word noises and long pauses. The verbatim transcripts also included annotations to indicate situations like repairs and mispronunciations, which were likewise removed.

Responses were scored from the clean transcripts according to conventional scoring rules outlined briefly above in 2.1.3 and in detail in Semel et al. (2003). These rules include different stopping points (i.e., ceilings) for certain subtests and various rules for allowing partial credit on certain items. Further score calibration on each subtest then took place in consultation with a licensed SLP to increase scoring consistency and accuracy. Normative scores were calculated using the participants’ ages at their first visit if they entered a new age bracket on the normative table during their participation. Study data were collected and managed using REDCap (Research Electronic Data Capture) (Harris et al., 2009), a secure, web-based application designed to support data capture for research studies. Clean data was processed through double-blind data entry, then evaluated for discrepancies utilizing REDCap electronic data capture tools. The process for one subtest, RS, differed in that clean transcripts were not available, so verbatim transcripts were manually reviewed (albeit without double-blind data entry) and edited to remove initial utterances (e.g., “um, I think it was …”) and repeated words, which are not counted as errors. For each of the subtests, item scores are summed to create a raw score. These raw scores are then converted to norm-referenced scaled scores based on age and using a conventional look-up table. The scaled scores have an average of 10 with a standard deviation of 3. Three of the subtests’ scaled scores are summed (FS, RS, and WS), and using another look-up table, an Expressive Language Index (ELI) composite score is generated. This is a standard score with a mean of 100 and a standard deviation of 15.

2.2 Computational Models

2.2.1 System Architecture

Leveraging deep learning approaches, we develop a computational model that includes a modular, cohesive scoring system capable of producing item-level test scores directly from raw transcripts. The entire transcript (response word, phrase, or sentence) is tokenized and encoded into a variable-length vector, which is fed into a BERT (Bidirectional Encoder Representations from Transformers). BERT is a DNN-based language representation model that has achieved state-of-the-art performance in numerous downstream tasks such as sentiment analysis (Xie et al., 2020) and question-answering systems (Qu et al., 2019). Recent studies, including Chen et al. (2020) and our previous work on the FS subtest (Wang et al., 2020),
have reported the superiority of such models in clinical tasks. The model has been fine-tuned to
our task as described in Section 2.2.2, and item-level scores are predicted for each subtest. Next, the
model summarizes predicted item-level scores according to the scoring rules described in Section 2.1.4
and sequentially computes raw score, scaled score, and ELI composite score across four subtests. A visual
walk-through of the scoring system architecture is presented in Figure 1.

2.2.2 Fine-Tuned BERT Classifier

The BERT model was originally trained on pairs of sentences to predict whether one sentence immediately
followed the other in its original context, a task known as next sentence prediction, as well as a masked
language modeling task in which the model uses the surrounding context to predict a word that has been
removed from a sentence. Internally, BERT captures contextual information using multiple blocks of
self-attention mechanisms that model global dependencies between the sequence of input and output words.
Trained on a large corpus of unlabeled sentences—3.3 billion words from the English Wikipedia combined
with BooksCorpus (Zhu et al., 2015)—language representations produced by BERT have been shown to be
effective in capturing both syntactic and semantic aspect of the language (Rogers et al., 2020; Ettinger,
2020). Similar to the original BERT model, our task also involves pairs of “sentences”: prompts and
responses. Accordingly, we form our input pairs to include the target word, phrase, or sentences alongside
each response to provide the model with the same information a scorer would use, annotated with special
BERT tokens to indicate breaks between prompt and response. For example, if an EV prompt was a picture
of a spatula, and the child responded “pancake turner,” the sequence would be provided to the model as
“[CLS] spatula [SEP] pancake turner [SEP].” Given the prompt-response pairs along with manual scores
(which serve as the gold standard), we develop four separate Multilayer Perceptron (MLP) classifiers by
replacing the original output layer of BERT with a softmax layer with $k$ nodes (classes) to classify each
stimulus response to EV, FS, RS, and WS subtests into scores of $\{0,1,2\}$, $\{0,1,2\}$, $\{0,1,2,3\}$, and $\{0,1\}$
respectively. The model is trained using categorical cross-entropy loss function via stochastic gradient
descent algorithm.

2.2.3 Evaluation Metrics

To evaluate the performance of item-level classifiers, we computed the precision, recall, and classification
accuracy for the item-level scores for each subtest. We further evaluated the predicted raw and scaled
scores across sub-tests, in addition to the predicted composite score, using mean absolute error (MAE)
between predicted and true scores. To validate results and establish their independence from our specific
data sets, and also to reduce the overfitting problem, we used cross-validation (CV) techniques Kohavi
(1995) in which the training, development, and test sets are rotated over the entire data set. With this
approach, the optimal parameters of the models were learned only from the training examples and never
from the examples used to test the model. Each comprehensive scoring evaluation was performed using
5-fold cross-validation, shuffling the entire data set 20 times; thus each result given is the average of 100
trials.

3 RESULTS

3.1 Baseline Models

As a baseline for EV and WS scoring models, we developed support vector machine (SVM) classifiers
(adopted from the Scikit-learn toolkit (Pedregosa et al., 2011)) trained on 300-dimensional vectors
representing the mean semantic encoding of all the words in the response using the pre-trained spacY
word2vec (Honnibal et al., 2020; Mikolov et al., 2013) model “en_core_web_Lg”, an embedding model
trained on a large corpus of web-based English text. For the RS baseline model, we trained a SVM classifier
on hand-crafted features that represent the correctness of a response. Inspired by our previous work (Gale
et al., 2020), we trained an SVM classifier with correctness features extracted using Levenshtein edit distance to compute the frequency of errors (insertions, deletions, substitutions), as well as the number of correct words. In a variation of our previous work on the FS task (Wang et al., 2020), we developed a MLP classifier as baseline for FS using a sentence-level BERT embedding for each response. In contrast to our new models, this baseline model does not fine-tune the weights within BERT. Instead, only the MLP classifier is trained using the static embedding calculated by the pre-trained BERT model.

3.2 Fine-Tuned BERT Model

We fine-tuned a variant of the original BERT model known as DistilBERT (Sanh et al., 2019) in our experiments. DistilBERT was specifically designed to reduce the number of parameters in the BERT model (66M versus 110M). Fewer parameters decreases the footprint of the model, and increases computation speed; more importantly, though, it decreases the risk of overfitting to the training data. We additionally added a L2 regularization term with the constant weight of 0.001 to further avoid overfitting. We trained the model with the learning rate of 0.00001 on batches of 16 response sequences at a time. A portion of the training data amounting to 10% of the entire data set was held out for validation. We fine-tuned the weights using a categorical cross-entropy loss, and for every iteration over the entire training set (“epoch”), we also calculated the loss on the validation set. When the validation loss failed to improve for four consecutive epochs, training was stopped, and the model was restored to the state with the best validation loss.

3.3 Model Performance

For the evaluation, we compared the performance of our proposed models to baseline models using the evaluation metrics described in Section 2.2.3. To explore the effect of extraneous words in transcripts, we examined our scoring models using our two input formats from Section 2.1.4: clean (the scored response as annotated by trained examiners) and verbatim (which includes examiner speech and other superfluous chatter). Table 3 reports the accuracy of the baseline classifiers and our proposed BERT-based classifiers on each of the four subtests, separately trained and tested on clean and verbatim transcripts. Results, in the form of macro averages of 100 trials, indicate that the fine-tuned BERT models outperform baselines in terms of classification accuracy, precision, and recall across all four subtests. As expected, the clean transcripts are more reliably scored than the verbatim transcripts, with the exception of the RS subtest, in which verbatim transcripts performed slightly better.

The fine-tuned BERT EV and WS models were the most accurate overall: EV had a 96% accuracy on verbatim transcripts, while WS had nearly 99% accuracy on clean transcripts. The FS scoring model showed the weakest performance, with the verbatim transcripts yielding 80% and 83% accuracy for the baseline and fine-tuned BERT models, respectively. This is not entirely surprising: FS responses are known to be difficult to manually score reliably, while the WS task is scored essentially on whether the child gives exactly the correct response.

The same patterns held for averaged MAE metrics. The fine-tuned BERT models outperformed baseline models in all cases, though as the scores were summed and normalized for raw and scaled scores, the difference in average MAE between models narrowed. Again, clean transcripts were more easily scored than their verbatim counterparts with the exception of RS, where difference is negligible. Full score MAE results are shown in Table 3. Figure 2 shows the distribution of scoring performance for each subtest, score tier, and transcript type over 100 trials. Note the smoothing effect as the score tiers progress, which is most pronounced in the FS and RS subtests: while the fine-tuned model clearly outperforms the baseline in item scores, the gap closes as scores are summed for a raw score, and yet again when the scaled scores are provided by the norm-reference table.
Aggregating the FS, RS, and WS scores to produce ELI composite scores, the baseline was substantially outperformed by the fine-tuned BERT model when using clean transcripts, with an MAE of 3.808 and 2.973 points respectively (on a scale of 45–155). Performance of the two models was more comparable when using verbatim transcripts, with an MAE of 3.939 and 3.706 points for the baseline and fine-tuned BERT models, respectively. These results are presented in Table 4, with the distribution over 100 trials also illustrated in Figure 2.

Lastly, we present Pearson correlations for predicted vs. true scores in Table 5. Again, the fine-tuned BERT model outperformed the baseline in all configurations. The highest correlation was found in RS raw scores, 99.0% and 98.9% for clean and verbose transcripts, respectively. The lowest correlation was in FS item scores: the baseline showed 75.1% and 69.2%, while the fine-tuned BERT model had 81.8% and 78.4% for clean and verbose transcripts, respectively. Overall, ELI composite scores correlated highly. The fine-tuned BERT model had correlations of 98.1% and 97.5% for clean and verbose transcripts, respectively.

4 DISCUSSION

In this paper, we describe a set of highly accurate computational models developed for scoring responses to several expressive language tasks for children. The models were combined into a comprehensive, multi-level scoring system based on conventional scoring methods. We automatically produced an item score for each subtest prompt, as well as a raw total score and a norm-referenced scaled score for each subtest. Three of the four subtests contribute to a composite score called the Expressive Language Index, which we also computed. The very high levels of accuracy of our results on several dimensions point to the utility and potential of automated computationally-driven methods of both administering and scoring expressive language tasks for pediatric developmental language evaluation.

4.1 Clinical significance

Our approach relies on a fine-tuned BERT modeling system which produced improved scores in all configurations. Even in the FS task, the most challenging subtest to score, the MAE of the scaled score, which ranges from 1 to 19, with a mean of 10 and standard deviation of 3, is just under 1 point. The CELF-4 manual considers a subtest scaled score below 7 points to be suggestive of clinical concern. An automated system that can produce scaled scores within about a point of the expert-derived score is likely to be clinically valuable. The system’s best accuracy for scaled score prediction was for the EV subtest, a confrontation naming task. Our new model has an MAE of well under half a point on verbatim transcripts, which demonstrates the potential of fully automated scoring of the EV task. Accuracy for this type of testing is particularly encouraging, as there are long used standalone language assessments with this focus (e.g. [Kaplan et al. (1983); Roth (2011); Martin and Brownell (2011)]). Automation of this type of language assessment could have great clinical relevance. Lastly, the ELI composite score is predicted within 4 points on the standard score scale of 45–155, generally well within the score’s confidence interval.

4.2 Comparison with prior work

The baseline models represent our previous best efforts in scoring each of the four subtests. The fine-tuned BERT methods presented here provide many natural advantages. Some of these advantages might explain the improvements over the baselines that we observe here, while others may hold promise for future advances in scoring, as well as deeper insight.

Our baseline RS model relied on explicitly engineered features and inherently discarded valuable information. For example, the Levenshtein method of error detection is unable to distinguish between an insertion and the superfluous chatter that distinguished our verbatim transcripts from the clean ones. The Levenshtein approach to scoring would be unable to distinguish between a common developmental...
error from a misrecognition in transcripts generated via automatic speech recognition. In addition, there may be latent clinically relevant information in responses that can be captured using contextualized language representation models like BERT but that would not easily be captured via rule-based scoring. The structure of BERT models extends well to introspective analysis, with techniques and tools for word-relationship heat maps and other such visualizations. (Kovaleva et al., 2019; Wu et al., 2020; Hanselowski and Gurevych, 2019).

For the EV and WS models, we used word2vec embedding models, which do not effectively capture contextual information and typically rely on sums or averages of individual word vectors. In addition, the pre-trained word2vec model we used in our baselines and the fixed DistilBERT model used with FS were both trained on texts written by adults and were not fine-tuned to our domain of child language. Fine-tuning the weights of the DistilBERT model allows us to take advantage of the broad representation abilities of a pre-trained language model while adapting it to the idiosyncratic language of our scoring task.

Another motivation for using the fine-tuned BERT model was the prospect of a unified approach to scoring different subtests. Although the content and scoring rules differ across the four subtests, our new system relies on a single architecture and differs only in terms of the data used to train the models. As a result, we benefit from a single pipeline for data preparation and feature engineering, and any improvements we make to the pipeline and architecture are more likely to apply to all subtests.

4.3 Future Work

Our vision for a fully automated scoring system starts with a spoken response. We intend to bring automatic speech recognition (ASR) into our system pipeline to automatically convert a participant’s recorded voice into written transcripts. The results presented here demonstrate how our fine-tuned BERT model adapts to the extraneous information present in verbatim transcripts. ASR transcripts contain the same extraneous information, and we expect a 10-20% word error rate on top of that. (Wu et al., 2019; Gale et al., 2019). Our previous experiments with the RS subtest (Gale et al., 2020) demonstrate the resilience of automated scoring when challenged with ASR transcripts, and we expect that our improved models may be even better equipped to adapt to ASR transcripts.

Our models could benefit from more (and more varied) training samples, but the time and effort required for test administration, transcription, and scoring makes it difficult to procure more data. Our previous work in (Wang et al., 2020) used machine translation technology to “translate” correct responses into partial- and no-credit responses, which was used as artificial data during training. That work showed how augmenting the training data with the artificial responses improved the performance of scoring models. We intend to incorporate this data augmentation technique into our new FS model, as well as adapt the technique to the other three subtests.

4.4 Conclusions

Even before the pandemic, many communities lacked regular access to clinicians and practitioners trained in assessing child language. The restrictions on travel and facility capacity that most of the world is now experiencing have further exacerbated this inequitable access to patient care. Methods of computerized administration and automatic scoring of language assessment instruments have the potential to reach underserved populations and to enable speech-language pathologists to devote more time to developing and applying interventions and treatments. Although it is certainly not the case that an automated system will ever be an adequate or comparable replacement for an expert clinician, technologies like the one proposed here can provide crucial support for these experts and for the schools and families that they serve.
CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

MA, JD, JB, JS, and EP contributed to the conception, design, and implementation of the data collection. RG designed, implemented, and evaluated the automated scoring system, with contributions from EP and YW. EP, JD, MA, and RG jointly wrote the manuscript, and all authors read and approved the submitted version.

FUNDING

Research reported in this publication was supported by the National Institute on Deafness and Other Communication Disorders and the National Institute on Aging of the National Institutes of Health under award numbers 5R01DC013996 and 5R21AG055749. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ACKNOWLEDGMENTS

We would like to acknowledge and thank the children and families who participated in this study. We would also like to thank our collaborative team of staff and volunteers at OSHU and BC.

REFERENCES

Boyle, J., McCartney, E., O’HARE, A., and Law, J. (2010). Intervention for mixed receptive–expressive language impairment: a review. Developmental Medicine & Child Neurology 52, 994–999

Chen, Q., Du, J., Kim, S., Wilbur, W. J., and Lu, Z. (2020). Deep learning with sentence embeddings pre-trained on biomedical corpora improves the performance of finding similar sentences in electronic medical records. BMC Medical Informatics and Decision Making 20, 1–10

Clegg, J., Hollis, C., Mawhood, L., and Rutter, M. (2005). Developmental language disorders—a follow-up in later adult life. cognitive, language and psychosocial outcomes. Journal of Child Psychology and Psychiatry 46, 128–149

Conti-Ramsden, G., Simkin, Z., and Botting, N. (2006). The prevalence of autistic spectrum disorders in adolescents with a history of specific language impairment (SLI). Journal of Child Psychology and Psychiatry 47, 621–628

Denman, D., Speyer, R., Munro, N., Pearce, W. M., Chen, Y.-W., and Cordier, R. (2017). Psychometric properties of language assessments for children aged 4–12 years: A systematic review. Frontiers in psychology 8, 1515

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Transactions of the Association for Computational Linguistics 8, 34–48. doi:10.1162/tacl_a_00298

Gale, R., Chen, L., Dolata, J., van Santen, J., and Asgari, M. (2019). Improving ASR Systems for Children with Autism and Language Impairment Using Domain-Focused DNN Transfer Techniques. In Proc. Interspeech 2019. 11–15. doi:10.21437/Interspeech.2019-3161

Gale, R., Dolata, J., Prud’hommeaux, E., van Santen, J., and Asgari, M. (2020). Automatic assessment of language ability in children with and without typical development. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC). 6111–6114
García Laborda, J. (2007). Introducing standardized ESL/EFL exams. *Language Learning & Technology* 11, 3–9

Grimm, A. and Schulz, P. (2014). Specific language impairment and early second language acquisition: the risk of over- and underdiagnosis. *Child Indicators Research* 7, 821–841

Hanselowski, A. and Gurevych, I. (2019). Analyzing Structures in the Semantic Vector Space: A Framework for Decomposing Word Embeddings. *arXiv:1912.10434 [cs]* ArXiv: 1912.10434

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., and Conde, J. G. (2009). Research electronic data capture (redcap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics* 42, 377 – 381. doi:https://doi.org/10.1016/j.jbi.2008.08.010

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python* (Zenodo). doi:10.5281/zenodo.1212303

Kaplan, E., Goodglass, H., and Weintraub, S. (1983). *The Boston Naming Test (2nd ed.)* (Lea & Febiger)

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International joint conference on artificial intelligence*. 1137–1143

Kovaleva, O., Romanov, A., Rogers, A., and Rumshisky, A. (2019). Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong, China: Association for Computational Linguistics), 4365–4374. doi:10.18653/v1/D19-1445

Marble-Flint, K. J., Strattman, K. H., and Schommer-Aikins, M. A. (2019). Comparing iPad® and paper assessments for children with ASD: An initial study. *Communication Disorders Quarterly* 40, 152–155

Marble-Flint, K. J., Strattman, K. H., and Schommer-Aikins, M. A. (2020). Comprehension scores among young neurotypical children and children with autism: Paper and iPad® storybooks. *Journal of Special Education Technology*, 0162643420924197

Martin, N. A. and Brownell, R. (2011). *Expressive one-word picture vocabulary test-4 (EOWPVT-4)* (Academic Therapy Publications)

[Dataset] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space

Noland, R. M. (2017). Intelligence testing using a tablet computer: Experiences with using Q-interactive. *Training and Education in Professional Psychology* 11, 156

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830

Qu, C., Yang, L., Qiu, M., Croft, W. B., Zhang, Y., and Iyyer, M. (2019). BERT with history answer embedding for conversational question answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1133–1136

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A Primer in BERTology: What we know about how BERT works. *arXiv:2002.12327 [cs]* ArXiv: 2002.12327

Rosenbaum, S. and Simon, P. (2016). *Speech and Language Disorders in Children: Implications for the Social Security Administration’s Supplemental Security Income Program.* (ERIC)

Roth, C. (2011). Boston naming test. In *Encyclopedia of Clinical Neuropsychology*, eds. J. S. Kreutzer, J. DeLuca, and B. Caplan (New York, NY: Springer New York). 430–433. doi:10.1007/978-0-387-79948-3_869

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR* abs/1910.01108
Semel, E., Wiig, E. H., and Secord, W. A. (2003). Clinical evaluation of language fundamentals, (CELF-4) 

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., and O’Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. Journal of speech, language, and hearing research 40, 1245–1260

Wang, Y., Prud’hommeaux, E., Asgari, M., and Dolata, J. (2020). Automated scoring of clinical expressive language evaluation tasks. In Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. 177–185

Wu, F., García, L. P., Povey, D., and Khudanpur, S. (2019). Advances in automatic speech recognition for child speech using factored time delay neural network. In Proceedings of Interspeech. 1–5

[Xdataset] Wu, Z., Chen, Y., Kao, B., and Liu, Q. (2020). Perturbed masking: Parameter-free probing for analyzing and interpreting BERT

Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. (2020). Unsupervised data augmentation for consistency training. Advances in Neural Information Processing Systems 33

Yew, S. G. K. and O’Kearney, R. (2013). Emotional and behavioural outcomes later in childhood and adolescence for children with specific language impairments: Meta-analyses of controlled prospective studies. Journal of Child Psychology and Psychiatry 54, 516–524

Zhu, Y., Kiros, R., Zemel, R. S., Salakhutdinov, R., Urtasun, R., Torralba, A., et al. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. CoRR abs/1506.06724

FIGURE CAPTIONS
Table 1. Baseline characteristics of participants

|                        | All (N=107) | TD (N=46) | ADHD (N=19) | DLD (N=22) | ASD (N=20) |
|------------------------|-------------|-----------|-------------|------------|------------|
| Female / Male          | 44% / 56%   | 57% / 43% | 53% / 47%   | 36% / 64%  | 15% / 85%  |
| Age in years, X (SD)   | 7.29 (1.05) | 7.03 (1.13)| 7.54 (1.18) | 7.54 (0.75)| 7.38 (0.97)|
| Race                   |             |           |             |            |            |
| Asian, N (%)           | 3 (3%)      |           | 3 (15%)     |            |            |
| Black / African American, N (%) | 3 (3%) | 1 (5%) | 2 (10%) | | |
| White / Caucasian, N (%) | 90 (84%) | 43 (93%) | 15 (79%) | 20 (91%) | 12 (60%) |
| More than one race, N (%) | 11 (10%) | 3 (7%)  | 3 (16%)    | 2 (9%)     | 3 (15%)    |
| Ethnicity              |             |           |             |            |            |
| Hispanic / Latino, N (%) | 10 (9%) | 2 (4%)  | 4 (21%)     | 1 (5%)     | 3 (15%)    |
| Not Hispanic / Latino, N (%) | 97 (91%) | 44 (96%) | 15 (79%)    | 21 (95%)   | 17 (85%)   |
| Language scores        |             |           |             |            |            |
| EV Scaled, X (SD)      | 12.48 (2.59)| 13.59 (1.90)| 12.89 (2.02)| 10.82 (2.99)| 11.35 (2.70)|
| FS Scaled, X (SD)      | 11.08 (3.67)| 12.70 (2.29)| 11.00 (2.36)| 10.09 (3.34)| 8.55 (5.58) |
| RS Scaled, X (SD)      | 10.62 (3.84)| 12.65 (2.70)| 10.42 (2.04)| 8.32 (3.72) | 8.65 (5.06) |
| WS Scaled, X (SD)      | 10.79 (3.25)| 12.46 (2.16)| 11.11 (2.45)| 9.09 (3.26) | 8.55 (3.89) |
| ELI Composite, X (SD)  | 105.32 (19.27)| 115.96 (10.61)| 105.26 (9.72)| 95.45 (18.60)| 91.75 (27.43)|

Table 2. Macro average (and standard deviation) of accuracy, precision, and recall measures for item-level scores. Experiments were conducted using 20 repeats of 5-fold classification, yielding 100 total experiments.

| Subtest | Transcripts | Accuracy |            | Precision |            | Recall |            |
|---------|-------------|----------|------------|-----------|------------|--------|------------|
|         |             | Baseline | BERT       | Baseline  | BERT       | Baseline | BERT       |
| EV      | Clean       | .979 (.002) | .984 (.003) | .979 (.001) | .984 (.003) | .979 (.002) | .984 (.003) |
|         | Verbatim    | .940 (.013) | .960 (.007) | .943 (.011) | .960 (.007) | .940 (.013) | .960 (.007) |
| FS      | Clean       | .8191 (.047) | .854 (.023) | .766 (.092) | .844 (.031) | .819 (.047) | .854 (.023) |
|         | Verbatim    | .8036 (.020) | .834 (.023) | .748 (.036) | .823 (.025) | .804 (.020) | .834 (.023) |
| RS      | Clean       | .836 (.018) | .866 (.014) | .846 (.012) | .864 (.014) | .836 (.018) | .866 (.014) |
|         | Verbatim    | .845 (.011) | .868 (.013) | .855 (.008) | .871 (.011) | .845 (.011) | .868 (.013) |
| WS      | Clean       | .969 (.015) | .987 (.006) | .969 (.014) | .987 (.006) | .969 (.015) | .987 (.006) |
|         | Verbatim    | .941 (.022) | .957 (.015) | .940 (.021) | .957 (.016) | .941 (.022) | .957 (.015) |
RESPONSE TRANSCRIPTS: For each of the subtests EV, FS, RS, and WS, we have transcripts (either clean or verbatim). Future versions automatic speech recognition (ASR) system.

TOKENIZATION & VECTORIZATION: The BERT tools provide a tokenizer to convert each item prompt/transcript pair into a vector. The numerical vectors have variable lengths, with special tokens at the beginning and to separate the prompt from the transcript.

ITEM SCORES: The deep learning architecture provides item-level predictions. The model was initialized with the weights from the pretrained DistilBERT model. We add our own output layer to predict probabilities for each score, e.g. with the FS subtest, we predict a probability for each of scores 0, 1, and 2. The score with the highest probability is selected.

RAW SCORES: As with paper-and-pencil scoring, the item scores are summed to provide raw scores.

SCALED SCORES: Using an age-specific norm-referenced table, the raw scores are converted into scaled scores on a range from 1-19.

COMPOSITE SCORE: The scaled scores for FS, RS, and WS are summed, and a norm reference is used to find the ELI composite.

Figure 1. Diagram of the scoring system architecture.
Table 3. Macro average (and standard deviation) of mean absolute error (MAE) for score estimation at the item score, raw score, and scaled score tiers. Results are presented for the baseline models from our previous work against the fine-tuned BERT models. Each of the EV, FS, RS, and WS subtests was scored automatically using clean and verbatim transcripts. Experiments were conducted using 20 repeats of 5-fold classification, yielding 100 total experiments.

| Subtest | Transcripts | Item Score MAE (std) | Raw Score MAE (std) | Scaled Score MAE (std) |
|---------|-------------|----------------------|---------------------|------------------------|
|         |             | Baseline | BERT | Scale | Baseline | BERT | Scale | Baseline | BERT | Scale |
| EV      | Clean       | 0.031 (0.003) | **0.026 (0.005)** | 0–2                  | 0.725 (0.078) | **0.610 (0.080)** | 0–54 | 0.261 (0.035) | **0.204 (0.079)** | 1–19 |
|         | Verbatim    | 0.093 (0.021) | **0.062 (0.013)** |                      | 1.771 (0.402) | **1.535 (0.223)** |              | 0.639 (0.165) | **0.415 (0.091)** |              |
| FS      | Clean       | 0.270 (0.091) | **0.194 (0.034)** | 0–2                  | 3.798 (2.736) | **3.072 (0.546)** | 0–48 | 1.437 (0.971) | **1.113 (0.144)** | 1–19 |
|         | Verbatim    | 0.301 (0.027) | **0.226 (0.025)** |                      | 3.639 (0.551) | **3.241 (0.363)** |              | 1.379 (0.213) | **1.275 (0.196)** |              |
| RS      | Clean       | 0.186 (0.023) | **0.154 (0.015)** | 0–3                  | 3.098 (0.299) | **2.407 (0.381)** | 0–96 | 0.661 (0.092) | **0.530 (0.107)** | 1–19 |
|         | Verbatim    | 0.173 (0.019) | **0.146 (0.015)** |                      | 2.582 (0.637) | **2.413 (0.473)** |              | 0.575 (0.155) | **0.504 (0.141)** |              |
| WS      | Clean       | 0.031 (0.015) | **0.013 (0.006)** | 0–1                  | 0.804 (0.367) | **0.355 (0.180)** | 0–32 | 0.514 (0.164) | **0.211 (0.085)** | 1–19 |
|         | Verbatim    | 0.059 (0.022) | **0.043 (0.015)** |                      | 1.309 (0.693) | **1.053 (0.377)** |              | 0.869 (0.406) | **0.657 (0.214)** |              |

Table 4. Mean absolute error (MAE) and standard deviation for Expressive Language Index (ELI) composite score with lookup table applied to scaled score estimation.

| Subtests | Transcripts | ELI Composite MAE (std) |
|----------|-------------|-------------------------|
|          |             | Baseline | BERT | Scale |
| FS, RS, WS| Clean       | 3.808 (1.785) | **2.973 (0.316)** | 45–155 |
|          | Verbatim    | 3.939 (1.253) | **3.706 (1.031)** |              |

Table 5. Correlation results for Item, Raw, and Scaled Score estimation.

| Subtest | Transcript | Item r² | Raw r² | Scaled r² | ELI r² |
|---------|------------|---------|--------|-----------|--------|
|         |            | Baseline | BERT | Baseline | BERT | Baseline | BERT | Baseline | BERT |
| EV      | Clean      | .963 (.009)** | .967 (.007)** | .984 (.008)** | .988 (.003)** | .973 (.008)** | .981 (.009)** |              |
|         | Verbose    | .889 (.026)** | **.925 (.014)** | .935 (.022)** | **.967 (.012)** | .894 (.053)** | **.958 (.013)** |              |
| FS      | Clean      | .751 (.049)** | **.818 (.054)** | .952 (.019)** | **.958 (.016)** | .851 (.248) | **.923 (.032)** |              |
|         | Verbose    | .692 (.053)** | **.784 (.044)** | .934 (.028)** | **.948 (.018)** | .880 (.053)** | **.894 (.044)** |              |
| RS      | Clean      | .935 (.011)** | **.944 (.007)** | .986 (.008)** | **.990 (.002)** | .976 (.070)** | **.981 (.003)** |              |
|         | Verbose    | .940 (.011)** | **.950 (.007)** | .988 (.006)** | **.989 (.004)** | .979 (.080)** | **.980 (.007)** | **.972 (.017)** | **.981 (.005)** |              |
| WS      | Clean      | .895 (.034)** | **.955 (.017)** | .948 (.045)** | **.989 (.005)** | .944 (.044)** | **.988 (.006)** |              |
|         | Verbose    | .793 (.050)** | **.853 (.038)** | .914 (.078)** | **.943 (.025)** | .886 (.113)** | **.944 (.025)** |              |

* (p < .01), ** (p < .001)
**Figure 2.** Mean (and standard deviation) of MAE results over 100 repeat experiments for item, raw, and scaled scores, as well as the ELI composite scores produced by baseline and fine-tuned BERT model. Results are shown for each subtest (EV, FS, RS, and WS), clean and verbatim transcripts.