The Reliability of Teacher Evaluations of Reading Skills of Primary School Children

Wieke Harmsen, Martijn Bentum, Ferdy Hubers, Roeland van Hout, Catia Cucchiarini & Helmer Strik
Radboud University Nijmegen, The Netherlands

Introduction

Reading proficiency in Dutch primary schools is assessed through word reading tasks: A pupil reads a list of words aloud, while a teacher scores the read words as correct or incorrect. This is a laborious task. An automatic system that evaluates a pupil's reading proficiency (decoding skills) could aid teachers in this process. To develop such a system, teacher evaluations (correct/incorrect) of pupils learning to read are needed. Because the system can only be as good as the data it was trained on, it is necessary to study the inter-rater reliability of such teacher evaluations. The current study aims to gain insight into the current practice of teachers with respect to the assessment of pupil's reading proficiency by addressing the following question:

To what extent are teacher evaluations reliable at the pupil level and at the word level?

Methodology

Dutch Automatic Reading Tutor (DART)

- **First graders (aged 6 – 7)** in Dutch primary schools were recorded while reading aloud **lists of 24 words** through DART (Bai et al., 2022)
- In total, **51 teachers** evaluated the words in the recordings as correct or incorrect.
- A subset of 6 recordings (**144 words in total**) were evaluated by all teachers to assess **inter-rater reliability**

Inter-rater reliability on the pupil level

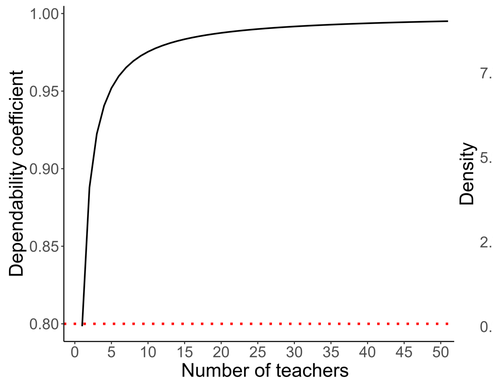
- The binary evaluations of each teacher for a single recording were aggregated into **a percentage correct**.
- Intraclass Correlation Coefficient, **ICC (2,1)**
- **Item-total correlations** for all teachers
- **Dependability coefficient** (generalizability theory) to assess number of teachers needed for reliable results on the pupil level

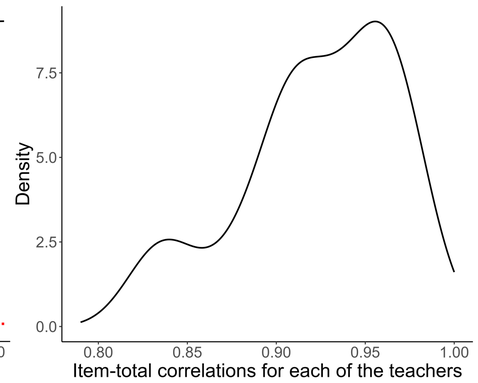
Inter-rater reliability on the word level

- Directly comparing the binary evaluations from different teachers for a specific word
- Because most words were rated as correct (72%), the **Matthews Correlation Coefficient** was used to assess agreement (Chicco et al., 2021)
- **Bootstrapping used** (sampling with replacement) to assess the number of teachers needed for reliable results on the word level

Pupil level results

Statistic	Value [95% CI]
ICC(2,1) (single teacher)	0.80 [0.60 – 0.96]
Mean item-total correlation	0.92 [0.90 – 0.93]





Reliability of judging reading proficiency at the pupil level is high

Conclusions

Teachers can reliably assess pupils' reading proficiency at the pupil level, but evaluations on the word level are less reliable

- Teachers probably use top-down information, but an automatic system to evaluate pupils' reading proficiency only works bottom-up
  - Thus, reliable teacher evaluations on the word level are required as training data
- More insight needed into the binary teacher evaluations on the word level (e.g. examine relation between teacher evaluations and type of reading errors)

Word level results

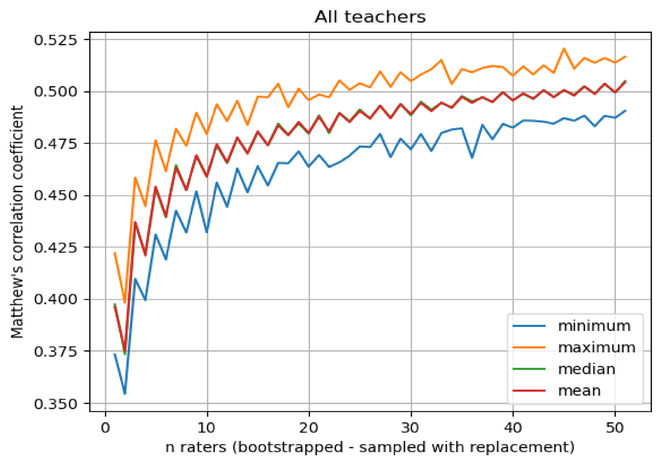
All teachers in agreement	# correct/incorrect words
41 words	40/1 words

- Teachers more often agree on correctly read words compared to incorrectly read words.

Matthews Correlation Coefficient (based on all teachers)

Mean [Median]	Min – Max
0.51 [0.74]	0.02 – 0.93

Bootstrapped Matthews Correlation Coefficients



Reliability of judging reading proficiency at word level is much lower

Acknowledgements
This work is part of the ASTLA project with project no. 06.20.TW.009, which is (partly) financed by the Dutch Research Council (NWO).

References
Bai, Y., Hubers, F., Cucchiarini, C., van Hout, R. & Strik, H. (2022). The Effects of Implicit and Explicit Feedback in an ASR-based Reading Tutor for Dutch First-graders. *Proceedings of Interspeech 2022*

Chicco, D., Warrens, M.J., & Jurman, G. (2021). The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary Classification Assessment. *IEEE Access,* vol. 9

Radboud University