ONLINE APPENDIX

Beyond one-size-fits-all: Differential effects of analogue and digital collaborative student-centred reading interventions in primary education

Franz Neuberger[1], Jana Heinz[2], Laura Eras[1] and Uta Hauck-Thum[3]

[1] German Youth Institute (DJI), Department of Children and Childcare, Munich, Germany
[2] Munich University of Applied Sciences, Faculty of Applied Social Sciences
[3] Ludwig Maximilian University of Munich (LMU), Faculty of Psychology and Education Sciences

Table A1: Descriptives by Intervention, MTP 1

Intervention	No Treatment					Analogue					Digital					
Variable	Mean	Sd	NotNA	Min	Max	Mean	Sd	NotNA	Min	Max	Mean	Sd	NotNA	Min	Max	Test
Reading skills	52	9.5	65	26	65	51	8.8	82	30	66	52	6.9	71	38	65	F= 0.311
Gender	75					89					82					X2= 4.069
... Girl	33	44%				44	49%				49	60%				
... Boy	42	56%				45	51%				33	40%				
Education	65					80					71					X2= 1.183
... No University	23	35%				25	31%				19	27%				
... University	42	65%				55	69%				52	73%				
Reading Motivation (Index)	3.2	0.85	72	1	4	3.2	0.8	88	1	4	3.2	0.69	79	2	4	F= 0.05
Language in the household	75					89					82					X2= 3.258
... mixed,foreign	35	47%				41	46%				48	59%				
... native	40	53%				48	54%				34	41%				
Usage of Digital Media (Index)	2.7	0.99	69	1	5	2.7	1.1	88	1	5	2.9	0.95	78	1	5	F= 0.894

Source: Digital Equity project, own calculations. For metric variables, the columns *Mean* and *Sd* each contain the mean and the standard deviation, for categorical variables *N* and the corresponding percentage values; no minimum or maximum is output here. The Test column reports an F (metric variables) or a χ^2 test (categorical variables), which indicates whether the distribution of the variable differs statistically significantly between the columns of the table. $*p < 0.1; **p < 0.05; ***p < 0.01$.

Model building

Node sizes

A critical consideration in tree-based models is the selection of the minimum node size, which determines the smallest allowed subgroup in the tree structure (Fokkema et al., 2018; Hothorn et al., 2006). This parameter substantially influences both model complexity and stability: too small node sizes may lead to overfitting and unstable trees, while too large node sizes might mask meaningful subgroup differences (Zeileis et al., 2008; Strobl et al., 2009). The challenge is particularly pronounced in longitudinal settings with nested data structures, where both the number of individuals and observations per individual need to be considered.

To empirically determine the optimal minimum node size for our LMERtree model, we conducted a bootstrap stability analysis with 100 replicates across different node sizes (30-90 observations). This resampling approach helps evaluate the robustness of tree structures across different subsets of the data (Strobl et al., 2009; Philipp et al., 2018). The stability analysis examined multiple metrics including the mean number of terminal nodes, their standard deviation, and the range of unique tree structures obtained across replicates. The analysis revealed that a minimum node size of 45 observations provided an optimal balance between model complexity and stability. Hence, this node size was chosen for the presentation of the results. At this node size, the trees averaged 6.71 terminal nodes (SD = 1.14) across bootstrap samples, with the number of terminal nodes ranging from 4 to 10. Smaller node sizes (e.g., 30 observations) resulted in more complex but less stable trees (mean = 8.91 nodes, SD = 1.74), while larger node sizes (e.g., 75-90 observations) produced overly simplified trees with only 2-6 terminal nodes. The selected node size of 45 observations, i.e. approximately 15 persons with 3 observations per node, representing approximately 10% of our person-level sample (n=184), ensure sufficient observations for stable parameter estimation within nodes while maintaining the model's ability to detect meaningful subgroup differences in our longitudinal data structure. See Table A2 below. The Figures A1 and A3 show a tree for the node size 30 and 60.

Overall, the trial runs with node sizes of 30 and 60 confirm the primary importance of motivation, media usage, and gender, even though these settings do not yield significant effects within the nodes (See Figures A2 and A4). However, e.g. the strong positive results for the digital intervention in Node 6, Figure A1 in particular indicate that in a larger sample there might be subgroups (unmotivated children of educated parents with low media usage) who also react positively to the digital intervention.

Table A2: Stability of Tree Sizes by Node Size

Node Size	Mean Nodes	SD Nodes	Min. Nodes	Max. Nodes	Unique Values
30	8.91	1.74	4.00	12.00	8.00
45	6.71	1.14	4.00	10.00	7.00
60	5.06	0.87	3.00	7.00	5.00
75	4.28	0.73	2.00	6.00	5.00
90	3.81	0.63	2.00	5.00	4.00

Results in Table A2 show descriptive statistics of the number of terminal nodes across bootstrap samples: mean nodes indicates the average number of terminal nodes, sd nodes represents the standard deviation, min nodes and max nodes show the range of terminal nodes, and unique values indicates the number of distinct tree sizes observed. For node size 45, which was selected for the final model, trees averaged 6.71 terminal nodes (SD = 1.14, range: 4-10 nodes) with 7 unique tree structures occurring across replicates, indicating a robust balance between stability and complexity.

Variable Selection

The stability analysis of our recursive partitioning model was further conducted using the stabletree function from the stablelearner R package (Philipp et al., 2018), which implements a bootstrap-based approach to assess variable selection stability. This function repeatedly fits trees to resampled versions of the original dataset, tracking which variables are selected as splitting variables across these bootstrap samples. By doing so, it quantifies the reliability and consistency of variable selection in the tree-building process. The stabletree function provides two key metrics, presented in table A3: the selection frequency of variables across bootstrap samples and their mean importance, with significance markers (*) indicating when these metrics exceed certain thresholds.

In our analysis, the stability results reveal a clear hierarchy of variable importance. Motivation emerges as the most robust predictor, demonstrating exceptional stability with a selection frequency of 98.8% across bootstrap samples and the highest mean importance value of 1.044. This strong performance in both metrics, each marked as statistically significant, underscores motivation's crucial role in the model's partitioning structure. Following motivation, both gender and media usage show moderate but noteworthy stability. Gender appears in 60.8% of the bootstrap samples with a mean importance of 0.618, while media usage is selected in 57.4% of samples with a slightly higher mean importance of 0.674. The significance markers indicate that gender's frequency and media usage's mean importance are particularly meaningful for the model structure.

In contrast, education and language demonstrated substantially lower stability, with selection frequencies of approximately 16% each. These low frequencies, coupled with their modest mean importance values, suggest these variables may not be reliable predictors for partitioning in our model. Most notably, the interaction term (inter) shows no stability whatsoever, with zero selection frequency and importance, strongly indicating it should be excluded from the final model specification.

These results provide clear guidance for model refinement, suggesting a focus on motivation as the primary partitioning variable, with potential consideration of gender and media usage as secondary factors. The stability analysis through stabletree helps ensure that our variable selection is not just based on a single model fit but is robust across multiple bootstrap samples, leading to a more reliable and parsimonious final model specification.

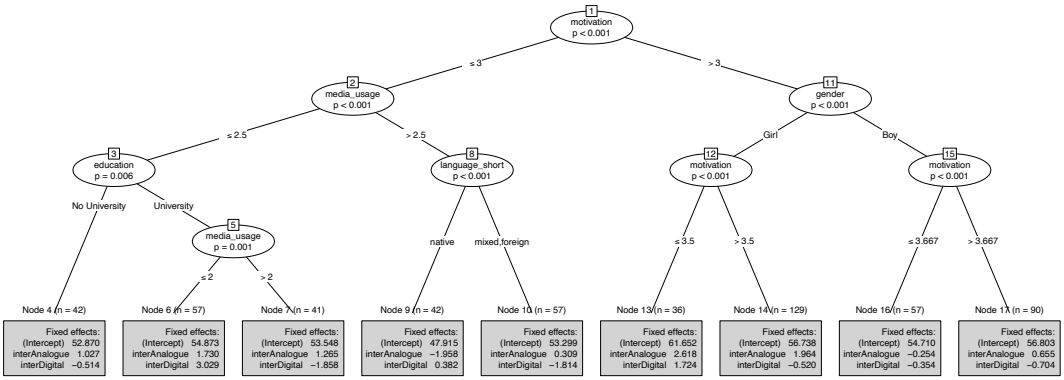
In light of these findings, our selected model with a minimum node size of 45 observations provides an optimal fit, as it exclusively incorporates the variables identified as important by the stablelearner analysis. As illustrated in the figure A1, reducing the node size (to 30) would lead to the inclusion of less stable splitting variables (such

Table A3: Variable Importance and Selection Stability

Variable	Selection Freq.	Sig.	Mean Importance	Sig.
Reading Motivation	0.99	*	1.04	**
Gender	0.61	*	0.62	*
Usage of Digital Media	0.57	*	0.67	**
Education	0.17		0.17	
Language HH: Mixed/Other	0.16		0.17	

Source: Digital Equity project, own calculations. Note: Sorting based on selection frequency. Selection frequency and mean importance based on 100 bootstrap replicates. Significance markers: $^{**}p < 0.01$; $^{*}p < 0.05$.

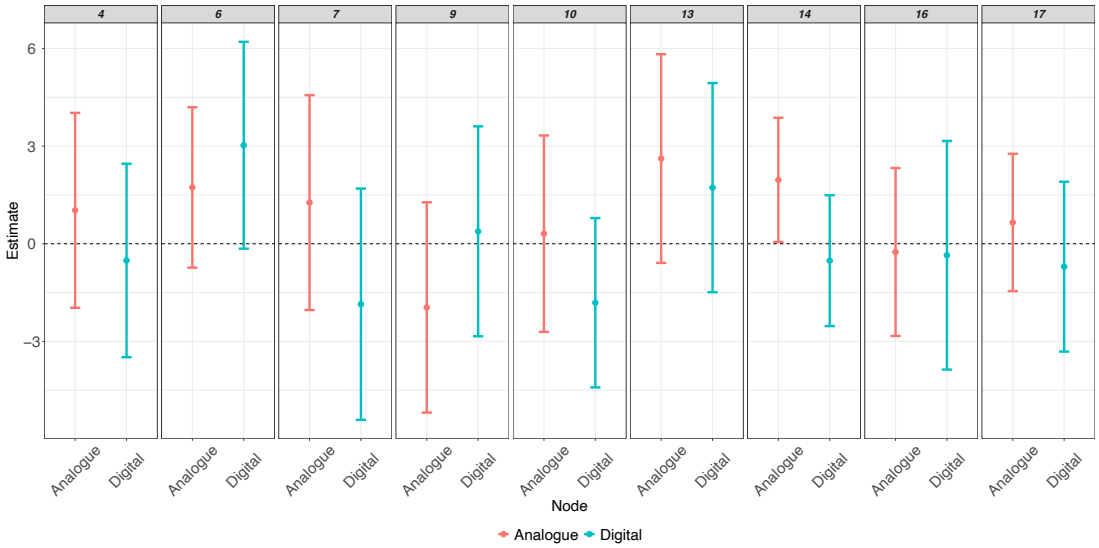
Figure A1: Node size 30: Effects of the intervention on reading skills in different subgroups



Source: Digital Equity project, own calculations. Splitting variables are gender, education, reading motivation, language in the household, usage of digital media. Minimal Node size restricted to 30.

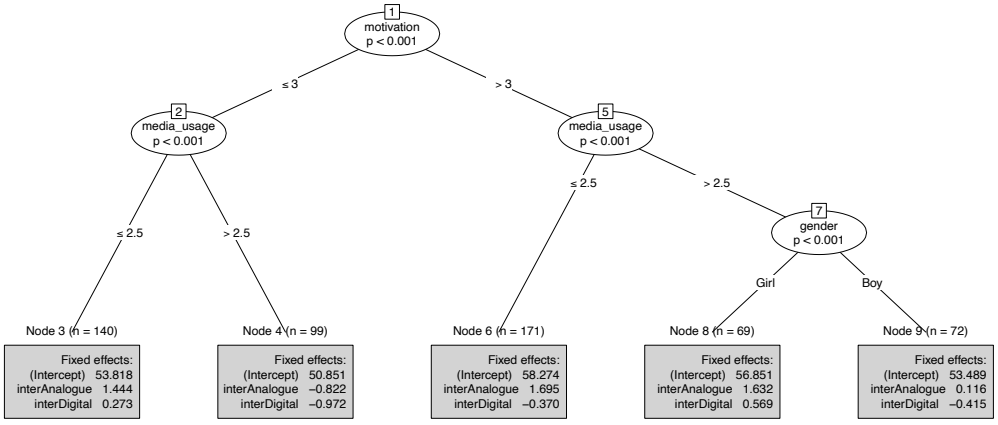
as education and language), which our stability analysis has shown to be unreliable predictors, thus potentially compromising the model's robustness and interpretability.

Figure A2: Effects from the Nodes from Fig. A1



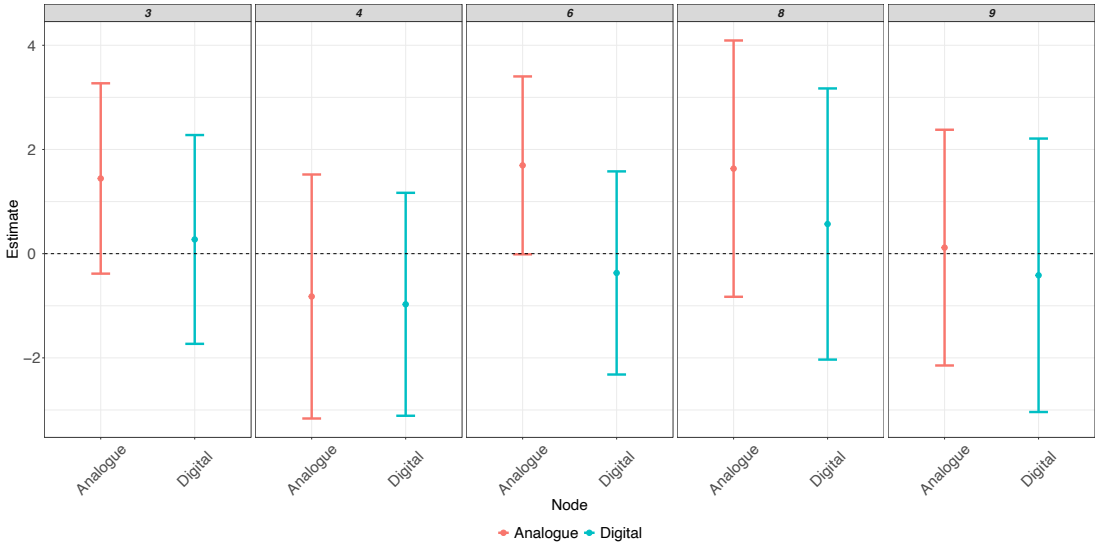
Source: Digital Equity project, own calculations. Effect sizes are from Endnodes from Fig. A1.

Figure A3: Node size 60: Effects of the intervention on reading skills in different subgroups



Source: Digital Equity project, own calculations. Splitting variables are gender, education, reading motivation, language in the household, usage of digital media. Minimal Node size restricted to 60.

Figure A4: Effects from the Nodes from Fig. A3



Source: Digital Equity project, own calculations. Effect sizes are from Endnodes from Fig. A3.

Literature

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. Behavior Research Methods, 50(5), 2016-2034.

Hothorn, T., Hornik, K., Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics, 15(3), 651-674. https://doi.org/10.1198/106186006X133933

Philipp, M., Zeileis, A., Strobl, C. (2018). A toolkit for stability assessment of tree-based learners. In B. Bischl, J. Bossek, D. Horn, T. Kerschke, L. Kotthoff (Eds.), Statistics Meets Machine Learning and Mathematical Optimization (pp. 315-325). Springer International Publishing.

Strobl, C., Malley, J., Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological Methods, 14(4), 323-348.

Zeileis, A., Hothorn, T., Hornik, K. (2008). Model-based recursive partitioning. Journal of Computational and Graphical Statistics, 17(2), 492-514.