

# CWTL RCT: Statistical analysis plan

---

Background .....	1
Flowchart .....	1
Dataset Creation and Cleaning .....	1
Summary of Demographic Variables.....	2
Missing Data Analysis and Multiple Imputation.....	2
Primary and Secondary Outcomes Effectiveness Analysis .....	3
Additional Analyses.....	4
References .....	5

## Background

War Child Holland has developed, along with partners, the Can't Wait to Learn (CWTL) programme, a tablet-based serious gaming educational programme. This statistical analysis plan (SAP) describe procedures for analysing data from a two-arm cluster randomized trial in Uganda. The trial compares the effectiveness of CWTL integrated into education as usual (EAU) with education as usual in increasing the two primary outcomes, numeracy and English reading competency, and secondary outcomes of children's psychological wellbeing and teachers' psychological wellbeing. Outcomes were evaluated at baseline (T0) and endline (T1); the period of programme implementation was 6 months.

Primary analyses will focus on difference in change in outcomes from baseline to endline between participants who received CWTL with EAU vs. EAU alone. The statistician conducting the analysis is blinded to the condition allocation.

## Flowchart

A flowchart of study participants will be created following CONSORT guidelines. The flow diagram will include the number of caregivers who consented to participate, the number of baseline assessments done with child, the number who completed the programme and reasons for not completing among those who did not, the number who completed the endline assessment and reasons for not completing assessments for those who did not, and the number included in analysis.

## Dataset Creation and Cleaning

1. A password protected tablet was used to collect data by independent research assistants. Kobo Toolbox software was used for the data collection. The data on the tablet was synchronized and uploaded first on the Kobo Toolbox Server. From this server it was downloaded and subsequently uploaded on the Sharepoint server of War Child Holland (WCH). All WCH laptops were encrypted and password protected. Data was checked for quality and validity throughout baseline and endline data collection by the research coordinator and researcher. School

attendance data was also collected by research assistants for the duration of programme implementation.

2. Data will be imported with R and a code for data cleaning will be created. Cleaning will include:
  - a) Checking for duplicate entries and incorrect IDs
  - b) \*Correcting errors in data entry that were noted during data collection
  - c) \*Checking all variables/measures for irregularities
    - i. Discrepancies or irregular data will be checked with field team
  - d) \*Response options for 'refuse', 'don't know,' 'not applicable' will be set to missing (.) in R
  - e) Variable names will be changed for easier identification if needed
    - i. A codebook for all variables will be created.

\*note that these steps are done separately for the baseline and endline data sets

3. A 'time' variable will be added to each of the databases
  - a) Baseline: time=0
  - b) Endline: time=1
4. Variable names across all databases will be identical where applicable except for the suffix denoting the time point (ID number will be the only variable not given a suffix)
  - c) Baseline: \_0
  - d) Endline: \_1
5. The databases will be merged on ID number. Because we specified in the previous step prefixes to variable names, we will be able to identify variables that were collected from each participant type in the new merged dataset. Two versions of the dataset will be created: a full WIDE and a full LONG version of the dataset.

## Summary of Demographic Variables

1. Using the WIDE dataset, participant baseline demographic data will be summarized stratified by treatment group. No inferential statistical tests (e.g., *t*-tests, chi-squared tests) with corresponding *p*-values will be conducted.
  - a. Continuous data (normally distributed): means, SDs
  - b. Continuous data (non-normally distributed): medians, IQRs
  - c. Categorical data: counts, frequencies, percentages
  - d. This analysis will populate Table 1 for the primary trial paper
2. Using the WIDE dataset, all primary and secondary outcome scales will be summarized for both time-points. Distribution of data and outliers for each scale will be examined.

## Missing Data Analysis and Multiple Imputation

We distinguish between missing data at item level and at outcome level (i.e. participants lost to follow up [LTFU]).

There will be no missing item values for numeracy or reading outcomes because of the way the data collection tools were constructed—if a participant could not answer a more basic question correctly, subsequent, harder questions are assumed to be incorrect (not missing). However, during the first round of baseline data collection, there was a glitch in the data collection form that did not allow assessors to use this skip logic for one question. Missing item values for the reading assessment item which measures the number of words read correctly in a passage (item name: 'RCP2') is possible in the baseline data due to data entry error resulting in impossible values. Missing data for this item will be

imputed using the answer to item 'RCP1', which measures the number of words read in the same passage in 60 seconds. This is because we know that the child can score at least as high as RCP1 (since RCP2 includes RCP1).

Missingness at outcome level are defined as missing sub scale scores at the endline due to dropout/LTFU. Logistic regression analyses will be used to explore whether demographic variables are associated with the dropout using a p-value of .05. If no variables are associated with dropout/LTFU, the missing data mechanism will be considered *missing completely at random (MCAR)*. If any variables are significantly associated with dropout/LTFU, the mechanism will be considered *missing at random (MAR)*. The percentage of missingness will be calculated and Cheema's (2014) recommendation will be followed to choose the order of the analyses. If the proportion of missing data is less than 15%, we will use multiple imputation in the main analysis (i.e. Intention to Treat analyses) and conduct complete case analysis to assess the sensitivity of our findings to the missing data. If the proportion of the sample lost to follow up is 15% or higher, the main analysis will be complete case analysis and multiple imputation will be used as a sensitivity analysis.

For the multiple imputation we will conduct multiple imputation with chained equations (MICE). This method allows the simultaneous imputation of all variables with missing data and allows us to specify a separate imputation equation for each variable. All variables from the follow-up timepoints will be imputed contingent on factors associated with dropout. We will follow a general rule of thumb (Royston & White, 2011; Bodner 2008) and impute  $m$  datasets approximately equivalent to the % of missing data. For example, if the missingness is 10% then we will impute approximately 10 datasets. Following guidelines by Simons, Rivero-Arias, Yu & Simon (2015), we will impute scale scores directly (instead of underlying individual items) given that the imputation of all scale individual items for all participant types can be computationally infeasible (i.e., likely to result in non-convergence of the imputation model) and because imputation of scale scores directly has been found to perform similarly to single-item imputation when missing data is primarily due to LTFU rather than item level missingness, which will likely be the case in this study. Following imputation of the abovementioned datasets, we will have 4 data files:

- Original WIDE data
- Original LONG data
- MI WIDE data
- MI LONG data

## Primary and Secondary Outcomes Effectiveness Analysis

Analyses will be conducted for each of the outcomes described using the MI LONG dataset. The statistical model is a two-level mixed effects regression model. In the cluster randomized controlled trial students were nested within schools and the schools were randomly assigned to the CWTL or EAU condition. Schools were the unit of randomization and the students were nested within schools, making this a two-level cluster randomized controlled trial. A cluster randomized controlled trial is more complicated than a single level trial since there is more than one level as there is variation on both the individual and the school level.

The data for a cluster randomized trial is presented in hierarchical form, with students nested within schools. At the level-1, or student-level model, the score on the outcome, i.e. sub topic difference score, endline minus baseline is predicted by the experimental condition (CWTL=1, EAU=0), and the student levels covariates gender and attendance. Level 2 is the school level. The mean outcome score, i.e. the

average difference outcome score of students within the school is predicted by the experimental condition.

Primary outcomes	# items	Description	Internal consistency
Letter knowledge (LID)	26	Give name/sound of all letters in alphabet	0.927
Phonemic awareness (PHO)	10	Initial sound identification	0.932
Reading: Fluency (RCP 1)	1	Number of words (out of 98) read 60 seconds.	---
Reading: Comprehension (COM)	5	1 story with 4 literal and 1 inferential question	0.84
Missing numbers (MIS)	10	Identify missing number from pattern	0.844
Timed addition (ADD)	25	120 seconds (timed); single, double and triple-digit numbers; incl. carrying, adding 3 numerals	0.966
Timed subtraction (SUB)	25	120 seconds (timed); single, double and triple-digit numbers; incl. borrowing	0.977
<b>Secondary outcomes</b>			
Total score, Stirling Children's Wellbeing Scale	12	Each item is scored 0-4 (never-always)	

Our two-tailed null hypothesis states there is no difference in the outcome between the average child in the two treatment conditions. Our alternative hypothesis states that there is a statistically significant difference between the average child in the two treatment conditions and that this difference is not readily attributable to chance. If the data are balanced, that is, the number of schools is equal for the two conditions and the number of clusters are equal, we can use the results of a *linear mixed regression* to test the effect of the intervention. The test statistic is an *F* statistic, which compares variance due the intervention effect to inter-school variance, controlled for covariates. If the null hypothesis is true, the *F* statistic follows a central *F* distribution with 1 degree of freedom for the numerator and *J*-2 degrees of freedom for the denominator. Under the central *F* distribution, we would expect the *F* statistic to be approximately 1. If the null hypothesis is false so that there is an intervention effect difference, the *F* statistic follows a non-central *F* distribution with 1 degree of freedom for the numerator and *J*-2 degrees of freedom for the denominator.

## Additional Analyses

**Fidelity** will be presented as an average score for the treatment group, with descriptive detail as appropriate. **Competency** of teachers will be included descriptively.

**Completer analysis:** A completer analysis will be conducted, where the main analysis is replicated on a subsample of participants defined as completers according to the following definition: children who have 80% school attendance or higher and have completed both baseline and endline assessments.

## Overview of data & sources

Data type	Description	Source
Attendance	Individual-level attendance data for all P3 children in all 30 schools	RAs collected this fortnightly using the class registers
Fidelity	Includes data on attendance, tablet management, session duration, teaching quality/blended learning	POs using observation forms to collect this data, aimed
Competency	Includes attitudes towards EdTech and knowledge of CWTL design and implementation. Developed as a quality assurance tool.	Self-reported by teacher training participants prior to the start of CWTL implementation and during the refresher training conducted at the start of Term 3.





## References

Bodner, Todd. (2008). What Improves with Increased Missing Data Imputations?. Structural Equation Modeling. 15. 651-675. 10.1080/10705510802339072.

Cheema, J. R. (2014). A Review of Missing Data Handling Methods in Education Research. Review of Educational Research, 84(4), 487–508. <https://doi.org/10.3102/0034654314532697>

Royston, P., & White, I. R. (2011). Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *Journal of Statistical Software*, 45(4), 1–20. <https://doi.org/10.18637/jss.v045.i04>

Simons, C. L., Rivero-Arias, O., Yu, L.-M., & Simon, J. (2015). Multiple imputation to deal with missing eq-5d-3l data: should we impute individual domains or the actual index? *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation - Official Journal of the International Society of Quality of Life Research*, 24(4), 805–815. <https://doi.org/10.1007/s11136-014-0837-y>

Prof. Dr. Mark Jordans	Dr. Nikhit D'Sa	Jasmine Turner	Dr Samantha Bouwmeester
<i>Co-Principal Investigator</i>	<i>Co-Principal Investigator</i>	<i>Co-investigator</i>	<i>Statistician</i>
Signature 	Signature 	Signature 	Signature 
Date: 16/12/2022	Date: 16/12/2022	Date: 16/12/2022	Date: 16/12/2022