Toddler mental development interventions: Can machine learning play a part?

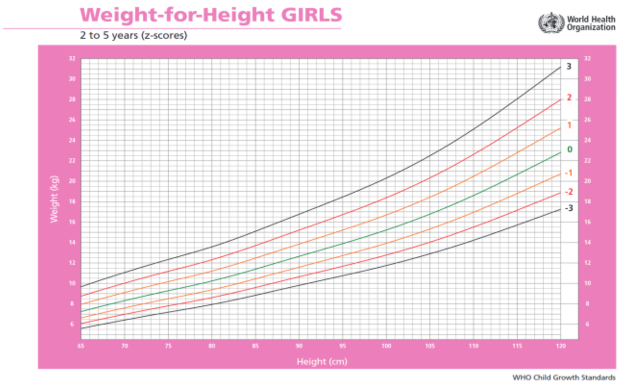
Akshat Gautam

Physical Development Monitoring

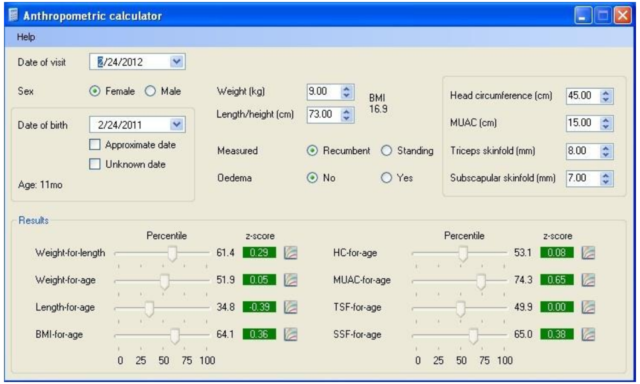
How is physical health measured in children?

- Physical health is monitored by measuring children's physical parameters such as height, weight, and head circumference
- There are standardized growth charts for children given by WHO.
- These measurements can be used to calculate z-scores, indicating how a child's measurements compare to typical values for their age and sex.

How is something like this done for mental development ?



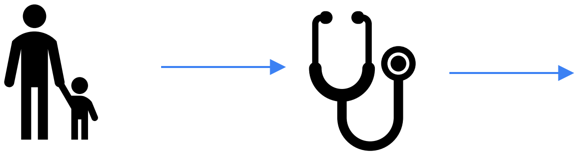
Height-Weight Chart [1]



WHO Anthro survey [2]

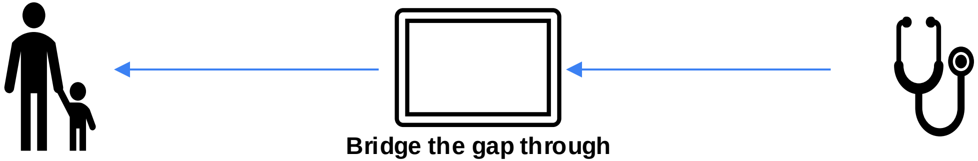
Gauging Mental Development

Goal : To measure and understand mental development in children.
Current Situation: Parents or caregivers observe atypical symptoms in a child's behavior, which necessitates a visit to the hospital.



- Hospital typically conduct psychometric tests
- These tests are often administered too late, after symptoms have already manifested.
- Results can be difficult to accept or interpret

Current project goal : Bring the hospital to children through tablet assessment and generate standardized mental development scores



Bridge the gap through tablet-based platform

Tablet Based Assessment

- Tablets contain a battery of tasks, each targeting different domains of development (e.g., social, motor).
- Each task generates data on backend which is used to generate features which represent child's performance

One of the task is **wheel task**, which targets social domain

Contribution 1: Extraction of feature from wheel task requires computer vision, and this feature is used into classification of children into NDD(Neurodevelopmental Disorder)/ TD (typically developing) . [Results]

However, there are more task which cover other domains of development like

Social:- Wheel Task(WT) , Preferential Looking Task (PLT), Button Task (BT)
Motor:- Motor Following Task (MFT), Colouring Task (CT), Bubble Popping Task (BPT)
Cognitive:- Delayed Gratification Task (DGT)

Understanding Psychometric Tests

What are these tests?

What are they:

- Psychometric tests are a standard and scientific method used to measure individuals' mental capabilities and behavioral style

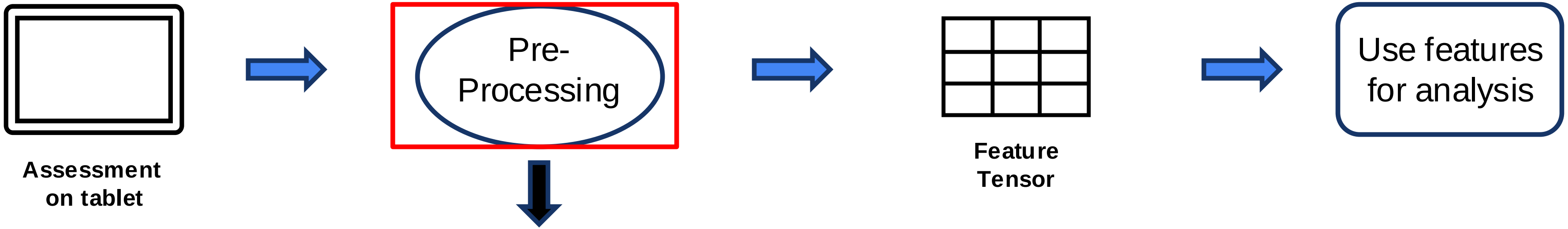
Why are they not used everywhere?

- Costly
- Need to be administered by trained professional in a specific setting
- Not available widely in Low-income countries

Overview

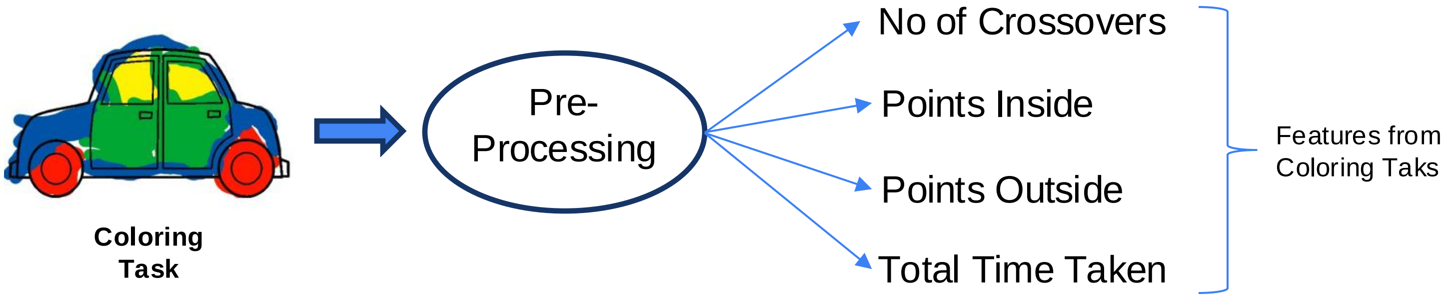
1. Introduction (done)
2. Pipeline
3. Understanding GMDS test (more detail)
4. Using features to predict GMDS scores
5. Using features to predict MDAT scores
6. Using IRT (item response theory) to generate scores
7. Future Work
8. Understanding tasks and feature extraction

Pipeline



Pre-Processing

The details of Pre-Processing to generate features will be covered later.
For now, we can assume each task can generate one or more features. For e.g.



What to do with these features?

Feature Tensor

Child 1	Coloring Task Feature	Wheel Task Feature	Button Task Feature	……….
Child 2	Coloring Task Feature	Wheel Task Feature	Button Task Feature	……….
Child 3	Coloring Task Feature	Wheel Task Feature	Button Task Feature	……….

Use features to predict developmental scores based on psychometric tests like MDAT/GMDS

Use features to generate developmental scores unsupervised (not dependent on psychometric tests)

Use features to classify into NDD/TD

(Done for wheel task only)

For this, it's important to understand psychometric tests mainly GMDS and MDAT

Griffith's Mental Development Scale (GMDS)

- Gold-standard tool in child development testing
- 0-6 years
- 321 items, 5 domains
  - **Foundations of learning**
  - **Language and communication**
  - **Eye and hand coordination**
  - **Personal-social-emotional**
  - **Gross motor**

Sample GMDS Results

ChildID	Domain A	Domain B	Domain C	Domain D	Domain E
MW-0113	33	43	44	53	47
IN-1653	49	51	52	56	60
IN-1682	31	43	46	45	46

Photo of GMDS test kit

Short demo of GMDS test

Testname	A	B
1.1	1	1
1.2	1	1
1.3	1	1
1.4	1	1
1.5	1	1
1.6	1	1
1.7	1	0
1.8	1	1
1.9	1	0
1.10	0	1
1.11	0	1
1.12	1	0
1.13	0	0
2.1	1	0
2.2	0	0
2.3	0	0
2.4	0	0
2.5	0	0
2.6	0	0
2.7	0	0
2.8	0	0
3.1	0	0
3.2	0	0
3.3	0	0
3.4	0	0
3.5	0	0
3.6	0	0
3.7	0	0
3.8	0	0
3.9	0	0
Scores	11	9

Age appropriate start point

Keep increasing level until 5 consecutive fails

Marks all levels after 5 fails as zero

Decrease levels until there are 5 consecutive successes.

Mark levels lower than these as also successes

Sum all successes to get a score

Repeat for all other 4 domains

How do we use these scores?

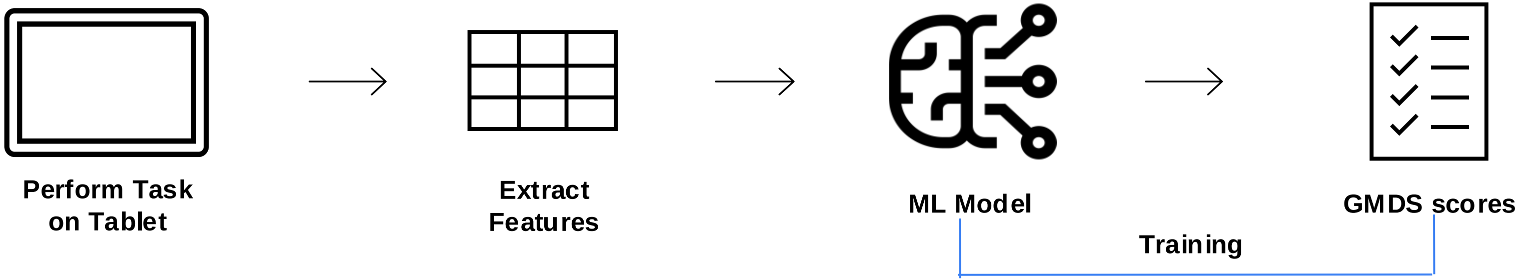
Motivation

Since the psychometric tests are not that easy to administer, can you get similar scores using some other method

Contribution

Use features generated from tablet-based tasks to generate developmental scores which are like psychometric tests

Pipeline



Setup

Training Data

384 data points (i.e. features and scores for 384 children)

56 features (54 features from 6 different tasks + Age,Gender)

Target label is GMDS scores across 5 domains

Training Setup

5-fold cross-validation (due to less data)

Metrics

R2 Score

Mean square error (MSE)

Mean absolute percentage error (MAPE)

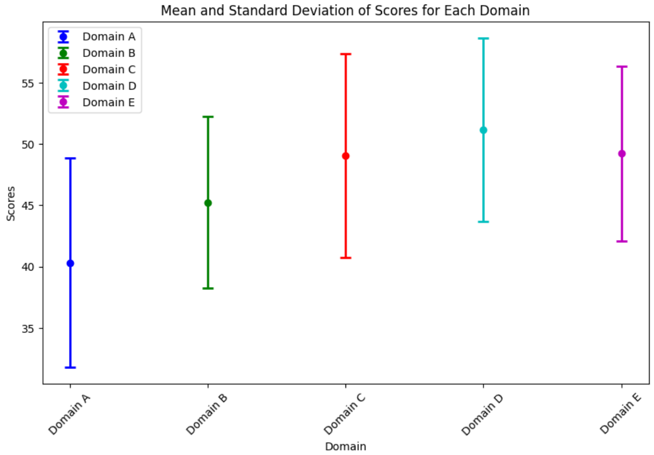
List of Models Used
• Linear Regression • Ridge Regression • Random Forest • Gradient Boosting • AdaBoost • Decision Tree • Support Vector Regression • KNN regressor • XGBoost

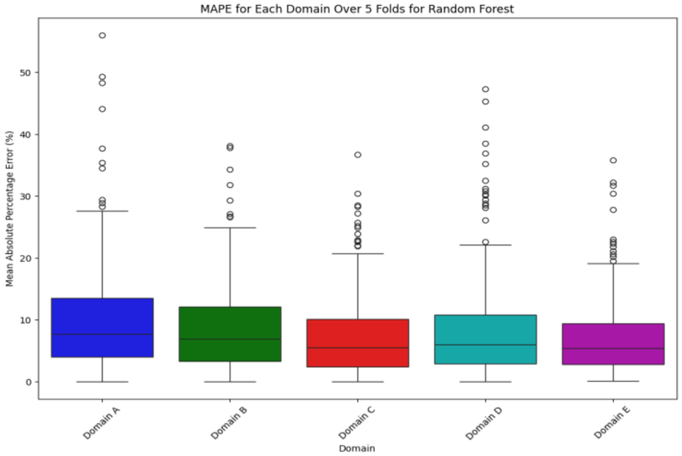
Results (1)

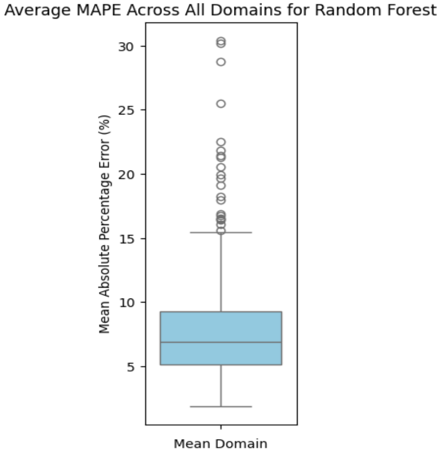
All the results are averaged over the 5 folds:-

Model	R2 Score	MSE	MAPE
Linear Regression	0.60	22.31	8.31%
Ridge Regression	0.62	21.56	8.19%
Random Forest	**0.64**	**20.12**	**7.88%**
Gradient Boosting	0.61	21.78	8.16%
AdaBoost	0.63	20.85	8.24%
Decision Tree	0.33	38.18	10.68%
Support Vector Regression	0.45	31.76	10.23%
KNN regressor	0.43	32.83	10.52%
XGBoost	0.57	24.42	8.73%

Results (2)

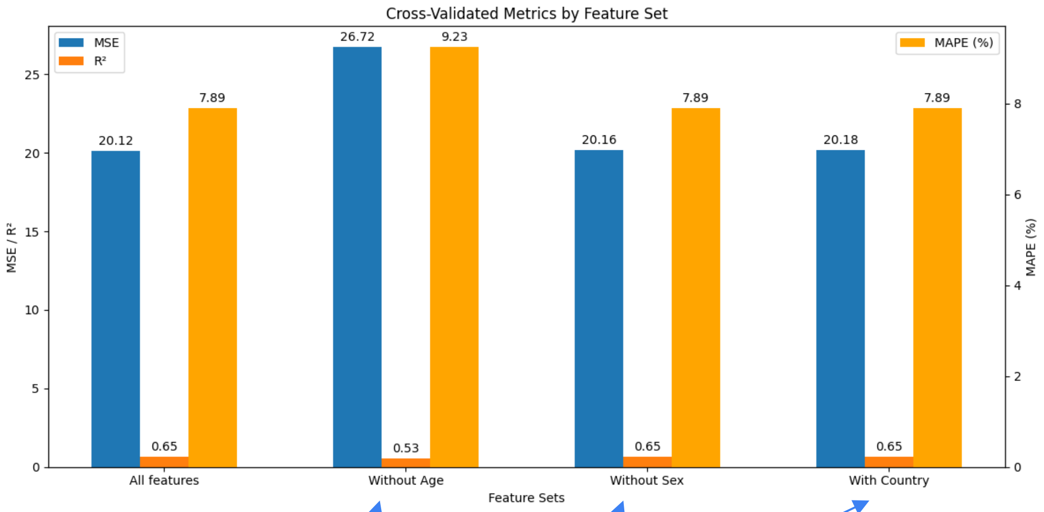






- First plot shows the mean and standard deviation of GMDS scores over the 5 domains
- Second plot shows the box plot for MAPE over the test samples for all the 5 domains
- Third plot shows the box plot for MAPE over the test samples averaged over the 5 domains
- **Model can predict GMDS scores with an error of 5-9% for almost half of the samples**

Results (3)



- Removing Age degrades performance of model as expected (GMDS scores highly correlated with age)
- Both removing Sex and adding Country as input to model doesn't change the performance

MDAT (Malawi Development Assessment Tool)

- Like GMDS but more culturally appropriate to low-income countries like Malawi
- Also, for age group 0-6 years
- 136 items across 4 domains
  - Gross Motor
  - Fine Motor
  - Language
  - Social

What's different from GMDS

- Task and Language more appropriate for countries like Malawi
- Done for every child unlike GMDS (more data point)

Sample MDAT Results

ChildID	Gross Motor	Fine Motor	Language	Social
MW-0113	48.44	51.60	49.95	45.23
IN-1653	50.37	48.13	53.32	52.79
IN-1682	48.40	47.15	51.56	48.07

Setup

Training Data

1459 data points (i.e. features and scores for 1459 children)

56 features (54 features from 6 different tasks + Age,Gender)

Target label is MDAT scores across 4 domains

Training Setup

5-fold cross-validation

Metrics

R2 Score

Mean square error (MSE)

Mean absolute percentage error (MAPE)

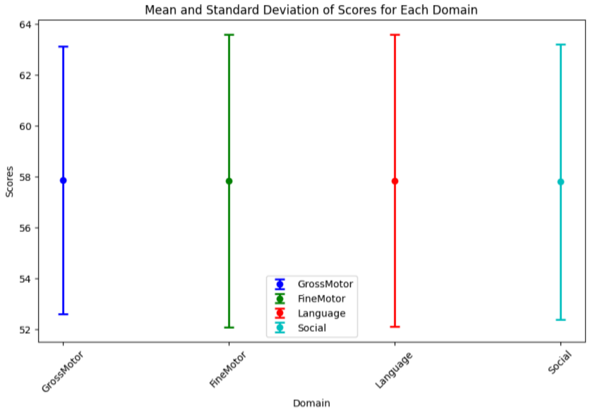
List of Models Used
• Linear Regression • Ridge Regression • Random Forest • Gradient Boosting • AdaBoost • Decision Tree • Support Vector Regression • KNN regressor • XGBoost • Neural Network

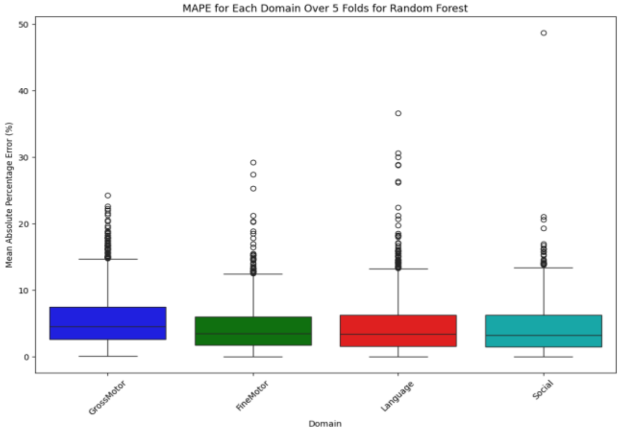
Results (1)

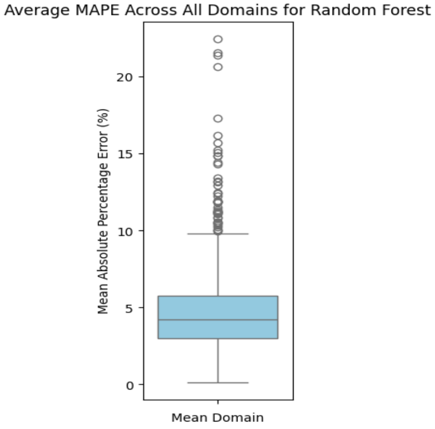
All the results are averaged over the 5 folds:-

Model	R2 Score	MSE	MAPE
Linear Regression	0.57	12.74	4.86%
Ridge Regression	0.58	12.66	4.85%
Random Forest	**0.60**	11.97	**4.61%**
Gradient Boosting	**0.60**	**11.95**	4.65%
AdaBoost	0.50	14.87	5.54%
Decision Tree	0.23	22.88	5.89%
Support Vector Regression	0.53	14.14	5.02%
KNN regressor	0.40	18.12	5.98%
XGBoost	0.55	13.31	4.75%

Results (2)







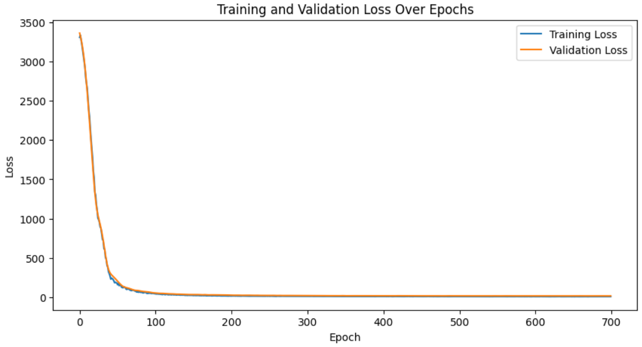
- First plot shows the mean and standard deviation of MDAT scores over the 4 domains
- Second plot shows the box plot for MAPE over the test samples for all the 4 domains
- Third plot shows the box plot for MAPE over the test samples averaged over the 4 domains
- **Model can predict MDAT scores with an error of 3-6% for almost half of the samples**

Results (3)

Since we have more data points, we can try to fit a neural network

Architecture

- The neural network consists of two fully connected layers: a 64-unit hidden layer and an output layer, both initialized with Xavier uniform initialization
- The hidden layer uses the Mish activation function, while the output layer uses ReLU.



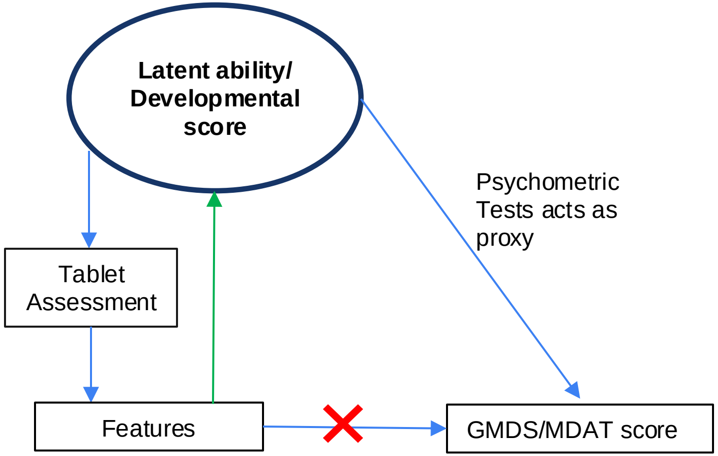
Epoch	MSE	R2 Score	MAPE
692	18.67	0.36	5%

Neural Network seems to not perform as good as other ML models.

- Tabular data
- Overfitting

Where are we now?

- Features extracted can be used to generate scores like GMDS
- Features extracted can be used to generate scores like MDAT
- Developmental scores could be generated independently



What's Next?

To generate developmental scores without relying on psychometric test

Item Response Theory (IRT)

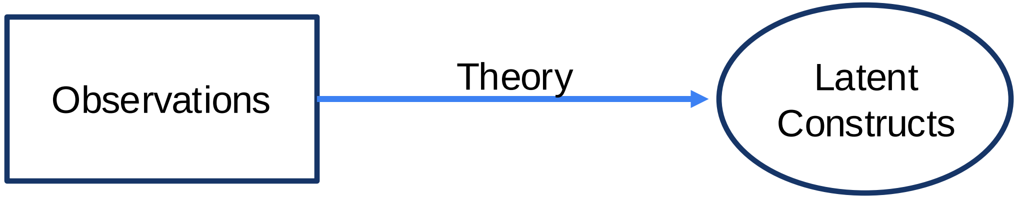
What is IRT?

- Information Response Theory is a theory of measurement, more precisely psychometric theory.
- Family of statistical models

What does it do?

IRT helps to map observations onto internal traits / states :-

- Test scores responses into knowledge / intelligence
- Questionnaire items into attitude / beliefs



More details

Measurement Tool

- Often a test/questionnaire consisting of several ‘items’
- Could be yes/no questions (could be task responses in our case!)

Measurement Theory

- Participant has an unobserved trait e.g. intelligence, knowledge, anger etc.
- Output of measurement tool is mapped to unobserved trait using some ‘scaling’

Summary

Questionnaires often involve mapping responses onto unobserved traits that are assumed to be continuous

But why not just add the responses to get the score

Every response is not same…
Answering 5 easy questions will results in different intelligence trait than answering 5 difficult question

Example

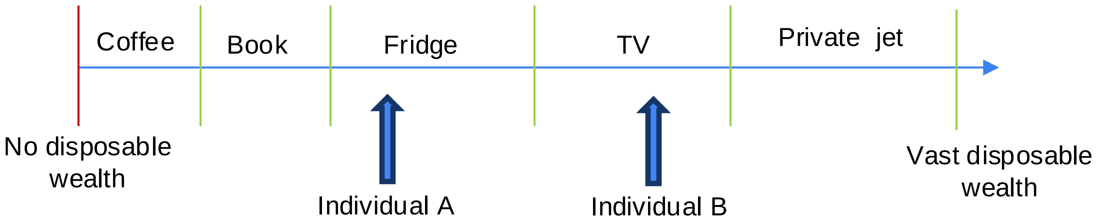
Problem Statement

To measure perceived disposable wealth of person

Questionnaire
• Can you afford a cup of coffee? • Can you afford a book? • Can you afford a fridge? • Can you afford a TV? • Can you afford a private jet?

Responses

Response	Individual A	Individual B
Coffee	1	1
Book	1	1
Fridge	0	1
TV	0	0
Private Jet	0	0



- We have two parameters to represent the choice that are: Item cost and participant wealth.
- Using these parameters, we want to move to probability space

Optimize Parameters

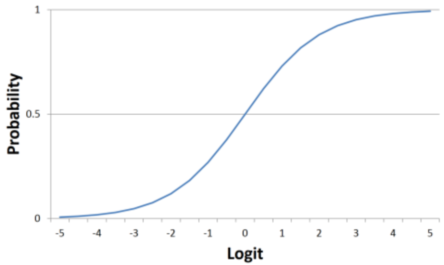
Probability	Individual A	Individual B
Coffee	0.75	0.95
Book	0.60	0.80
Fridge	0.40	0.60
TV	0.10	0.40
Private Jet	0	0.15

Moving to Probability Space : Rasch/1-Parameter model

$$Logit_{person,item} = \underbrace{Wealth_{person} - Cost_{item}}_{\text{Not between [0,1]}}$$

Thus, for mapping values to [0,1],

$$Logit = \ln(\frac{Pr}{1 - Pr})$$



So, in general

$$Y_{ij} = \theta_j - b_i$$

Where, Y_{ij}= Logit of Response by person j for item i,

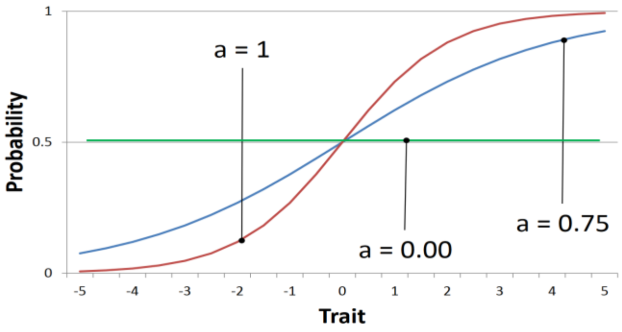
θ_j= **Trait** of person j**,**

b_i= Difficulty of item i

Some other models

2 Parameter Model
$Y_{ij} = a_i\theta_j \; - b_i$
Y_{ij}= Logit of Response by person j for item i,
a_i= Discrimination of item I,
θ_j= **Trait** of person j**,**
b_i= Difficulty of item i

Same difficulty, different discriminations



In STREAM, if we consider tasks metric as "items" in questionnaire, the responses would not be in binary. For e.g., for coloring tasks :-

Task metric / Items	Features / Responses
Points Inside	636
Points Outside	1595
Crossovers	63
Time Taken	88216

Adapting for STREAM data

What if we remove the logit link from the equation earlier :-

$$Y_{ij} = \theta_j - b_i$$

Where, Y_{ij}=Feature of child j for the task metric i ,

θ_j= Ability of the child j,

b_i= Difficulty of task metric j

Here the Y_{ij} would be continuous, this is also known as random intercept mixed effect regression model and it is very similar to Rasch/ 1-parameter model.

Fixed effect : Task metrics as they would be same across the children
Random effect: Each child’s ability that is development on those tasks

Setup for START data

Looking at the equations for just one task feature,

$$
\begin{aligned}
Y_{points\ inside,child\ 1} &= \theta_{child\ 1} - b_{points\ inside} \\
Y_{points\ inside,child\ 2} &= \theta_{child\ 2} - b_{points\ inside} \\
Y_{points\ inside,child\ 3} &= \theta_{child\ 3} - b_{points\ inside} \\
&\ \ \cdots\cdots
\end{aligned}
$$

Fixed effect

Random effect

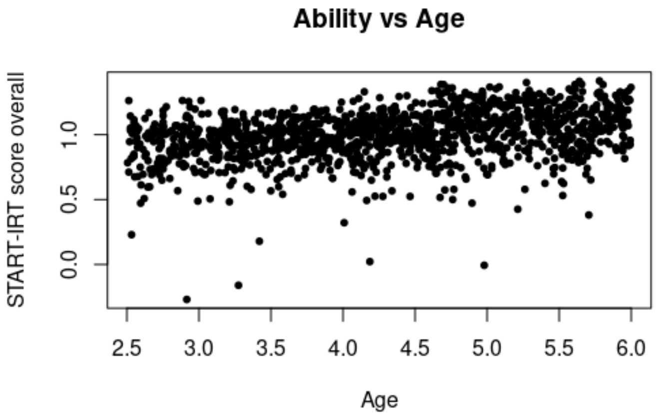
So finally, the final equation would be,

$$Feature_{\text{task},child} = \theta_{child} - b_{task}$$

However, we need to be careful that in this case, higher the value of feature we expect ability to be also higher.

Which is not always the case for e.g. we expect the crossovers to be lower for a child with higher development

Results(1)



r = 0.34

Correlation of IRT score with	r
MDAT total score	0.27
MDAT Gross motor	0.22
MDAT Fine motor	0.24
MDAT Language	0.24
MDAT Social	0.25

Correlating IRT scores with MDAT

Relatively low correlation with age and MDAT scores, the scores need to be improved

Improvements/Future Work

Why is correlation not good?

- IRT assumes monocity, in our case it means as ability increases the task feature should also increase
  - This may not be necessarily true since they are hand crafted features, for e.g. number of crossovers increases as age of children increases which is not expected

- IRT assumes local independence- responses given to the separate items in a test are mutually independent given a certain level of ability (multiple features are extracted from same task)

- We are using all features to predict a single score, could bin features into different domains and generate multiple scores like -> social, motor …

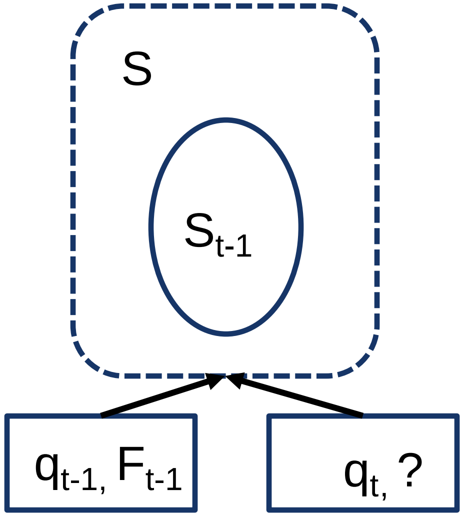
Future Work Motivation

- We are not including the fact that the child is performing tasks in a particular order, and in a single sitting
- Each feature may require mastery in multiple areas (social, motor, fine motor etc.), however we may not know the areas corresponding to each feature
- We are fitting a **linear** mixed effect regression model

Knowledge State

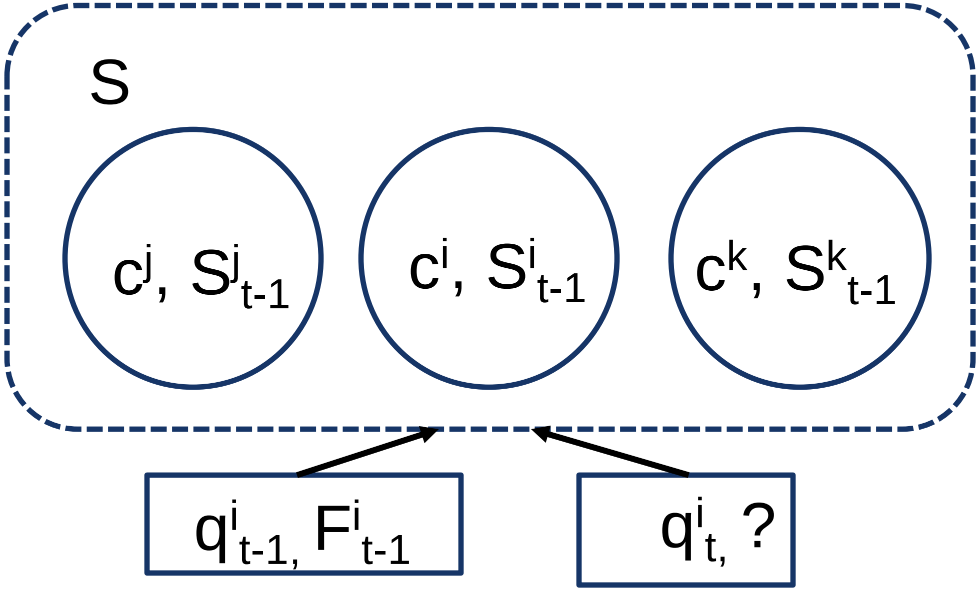
- Given a child's previous task attempts X= {x1,x2, … xt-1}, our goal is to predict the feature (say number of crossovers) that child will achieve in the current task
  - ➢ Each input $x_t = (q_t, F_t)$ is a tuple containing task q_t, and its feature F_t which is computed from the tablet data

- The information of previous attempted tasks is condensed into a latent knowledge state S={s1,s2, … st-1}
  - ➢ For example, if our previous method incorrectly predicts a feature F_t, our goal is to update the model and the knowledge state, thus improving our understanding of the child as she attempts task over time



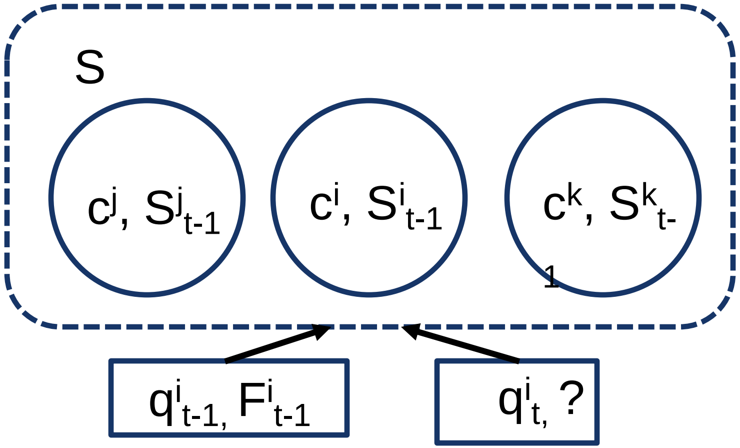
Concepts

We want to have multiple concepts for each state i.e c1,c2….cn



States and concepts

- Combine knowledge state and concept in a memory augmented neural net paradigm
- Training
  - Learn static matrix (key) for storing concepts associated with each task independent of child
  - Learn matrix (value) for storing student's knowledge state in each concept
- Inference
  - Update the value matrix as child completes task
  - Final score after all tasks are completed

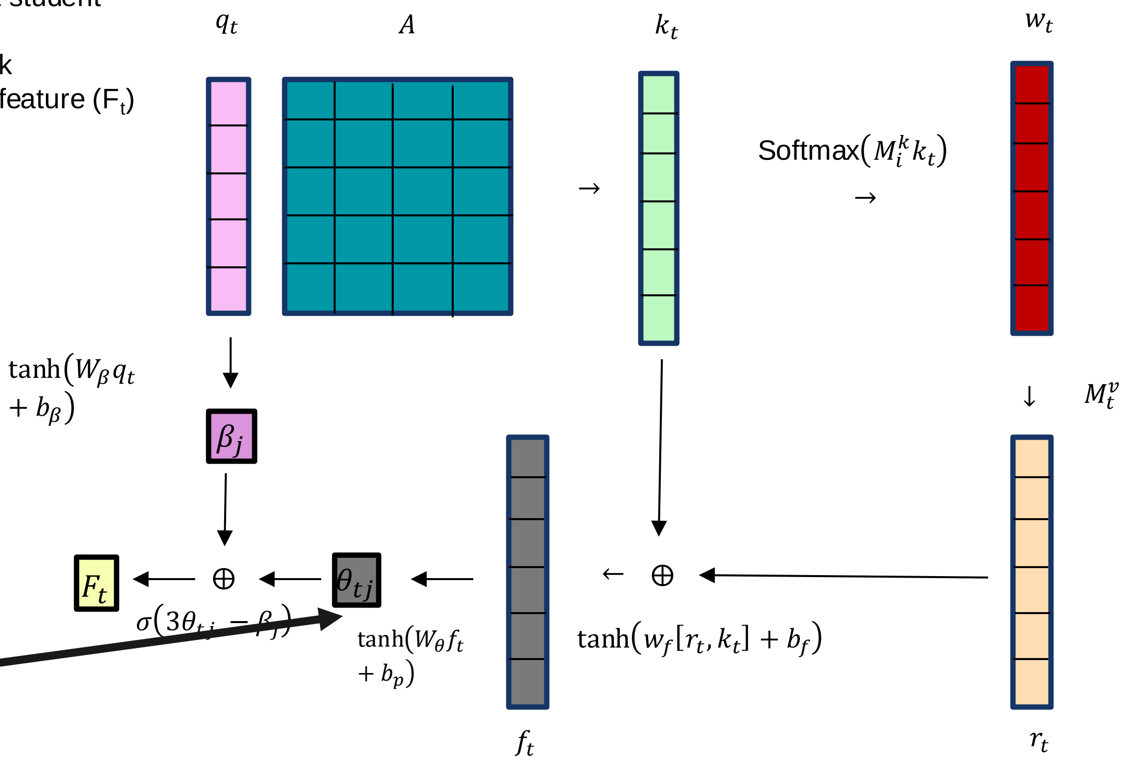


Maintains a knowledge state for each concept simultaneously and all states constitute the "knowledge" of a child

Method 3: Deep IRT

- $M^v \in \mathbb{R}^{N \times d_v}$: Value memory matrix (skill states)
- $M^{\mathrm{k}} \in \mathbb{R}^{N \times d_k}$: Key memory matrix (latent abilities)
- $A \in \mathbb{R}^{d_k \times Q}$: Ability Components Embedding matrix
- $k_t \in \mathbb{R}^{d_k}$: Embedding vector (key)
- $v_t \in \mathbb{R}^{d_v}$: Response Embedding vector
- $e_t \in \mathbb{R}^{d_v}$: Response erase vector
- $\mathbf{B} \in \mathbb{R}^{Q \times d_v}$: Ability Components response embedding matrix

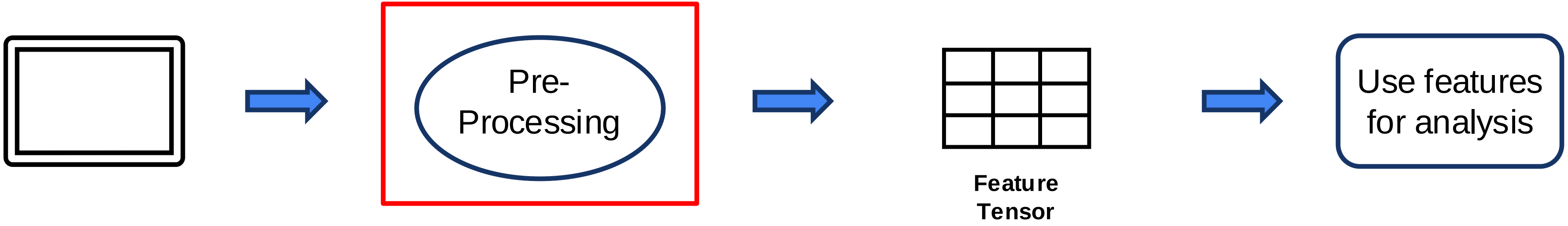
- IRT Module is built on top of key-value network
  - Instead of predicting probability, we predict student ability using f_t
- Task difficulty is calculated using another network
- Ability and difficulty is combined to predict task feature (F_t) [reminiscent of Item Response Theory]



Ability is our developmental score

Back to feature extraction

Now we know how features extracted can be used to generate developmental score. Data from tablet is stored at backend, which is processed to generate these features



There are separate ways to process each of the STREAM tasks, the tasks can be broadly divided into two areas:

- Light Data -> Involves processing excel files generated from backend
- Heavy Data -> Involves processing video files stored at backend

Wheel Task (heavy data)

Task Description

A black and white wheel appears on the screen, children are instructed to watch the wheel, while their video is recorded on the tablet

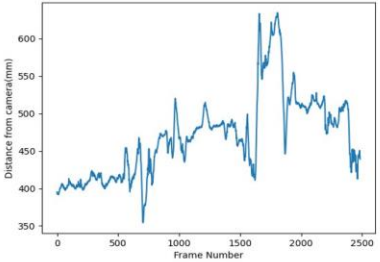
Feature description

Get the distance of face from the camera using a video

Not cover the distance extraction method for now

Black and white wheel

Video recorded from front camera

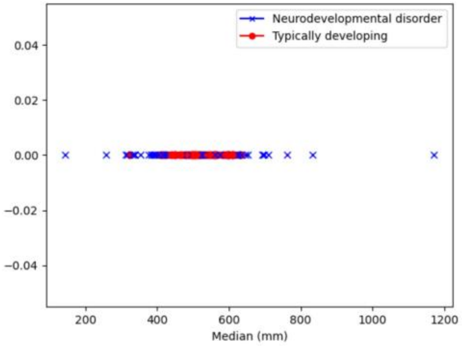


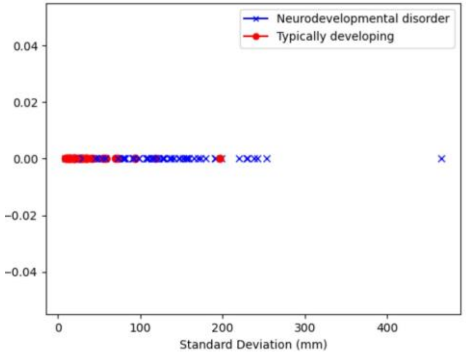
Distance vs Frame no

[Get Back](#)

Extracting feature

Distance signal can either be directly used or some features may be extracted from it. Through trials we found median and standard deviation to be effective in classification





- Medians of Neurodevelopmental Disorder(NDD) tend to deviate to more extreme values.
- Typically developing (TD) participants tend to display lower standard deviations.

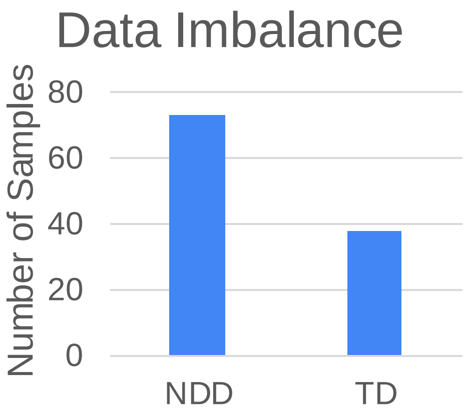
Classification

Input
For each child we have two features as input, median and standard deviation in mm

Training Setup
Models used Random Forest, Logistic Regression and SVM
Total data points **111**, 5-fold CV , accuracy over 5 test folds

Result and Discussion
Logistic regression showed highest accuracy of 0.81
F1 score 0.74
F1 score may be more important in our case due to high imbalance

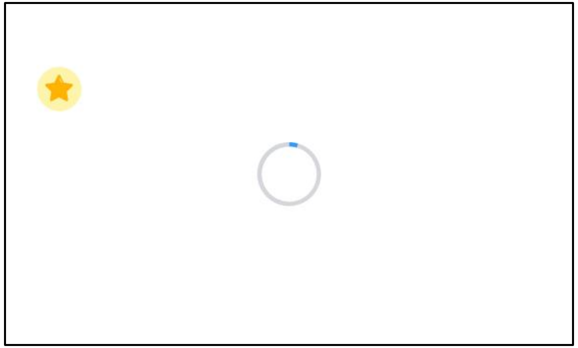
Algorithm	Accuracy(%)	F1 Score
Random Forest	78.46	0.67
Logistic Regression	81.23	0.74
SVM	73.07	0.55



Delayed Gratification Task

Task Description

A star appears on screen. Child is told to wait for some time to get all three stars.



Feature Description

1. Proportion time spent delaying gratification
2. Proportion of frames child's face visible

Start time and end time are read through the excel files. Total task time is 180 s

Proportion Time = $\frac{End\ Time - Start\ Time}{180}$

Medipipe face mesh is used to detect if a face is present or not in the frame. If more than one faces are present, then that frame is ignored.

Proportion face = $\frac{No\ of\ frames\ with\ a\ face}{Total\ No\ of\ frames}$

Summary

- The data stored in the backend from tablet assessments can be converted into relevant features.
- These features can be used for classification into NDD/TD.
- These features can be used to generate scores under the supervision of MDAT/GMDS scores.
- Item Response Theory could be used to generate developmental scores in an unsupervised setting.

Acknowledgements

- My advisor: Prof. Sharat Chandran
- Prof. Bhismadev Chakrabarti and the whole STREAM team
- Shubham (especially for distance work ,experiments ..)

Thank You!

References

[1] https://www.who.int/tools/child-growth-standards/standards/weight-for-length-height
[2] https://www.who.int/tools/child-growth-standards/software