

Affective computing for understanding neurological child development

Akshat Gautam
Prof. Sharat Chandran

Physical Development Monitoring

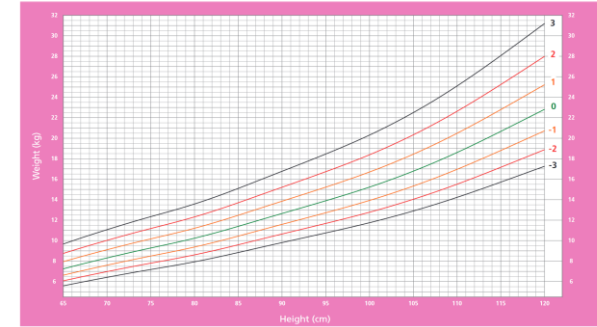
How is physical health measured in children?

- There are standardized growth charts for children given by WHO.
- Physical health is monitored by measuring children's physical parameters such as height, weight, and head circumference
- 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 ?

Weight-for-Height GIRLS

2 to 5 years (z-scores)



WHO Child Growth Standards

Height-Weight Chart [1]

Anthropometric calculator

Help

Date of visit: 8/24/2012

Sex: ☒ Female ☐ Male

Date of birth: 2/24/2011

☐ Approximate date
☐ Unknown date

Age: 11mo

Weight (kg): 9.00 BMI: 16.9

Length/height (cm): 73.00

Measured: ☒ Recumbent ☐ Standing

Oedema: ☒ No ☐ Yes

Head circumference (cm): 45.00

MUAC (cm): 15.00

Triceps skinfold (mm): 8.00

Subscapular skinfold (mm): 7.00

Results

| Measurement | Percentile | z-score |
|-------------------|------------|---------|
| Weight-for-length | 61.4 | 0.29 |
| Weight-for-age | 51.9 | 0.05 |
| Length-for-age | 34.8 | -0.39 |
| BMI-for-age | 64.1 | 0.36 |
| HC-for-age | 53.1 | 0.08 |
| MUAC-for-age | 74.3 | 0.65 |
| TSF-for-age | 49.9 | 0.00 |
| SSF-for-age | 65.0 | 0.38 |

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 ASD(Autism spectrum Disorder)/ TD (typically developing) . [\[Results Later\]](#)

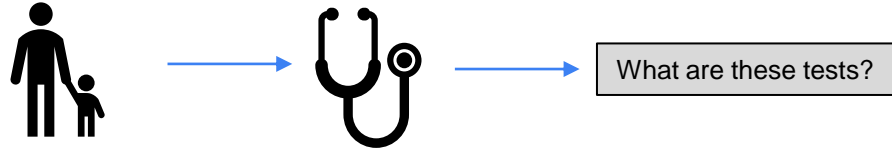
However, there are more task which cover other domains of development like

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 they:

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

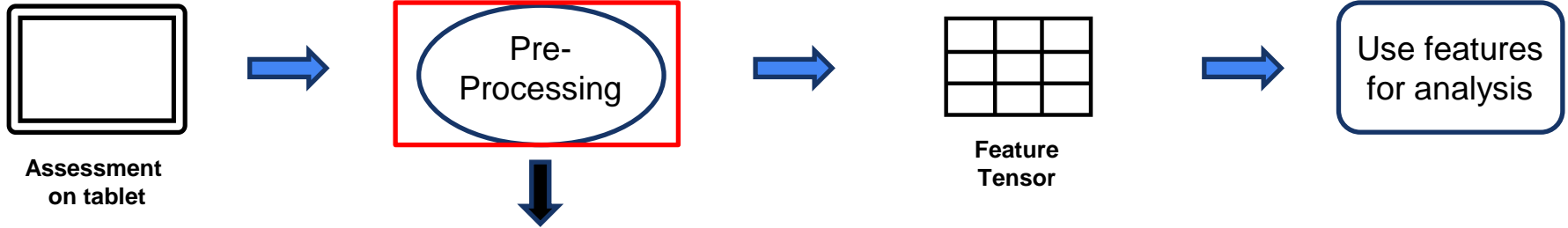
Why are they not used everywhere?

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

Overview

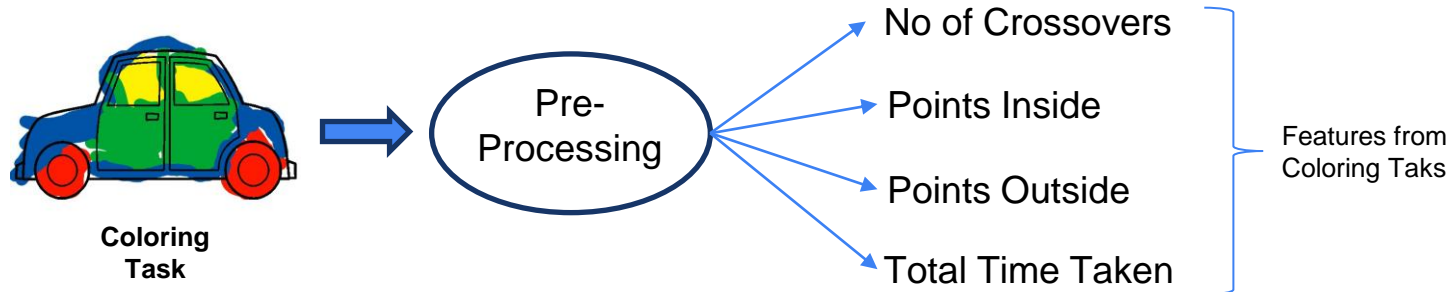
1. Introduction (done)
2. Pipeline
3. Understanding GMDS test (more detail)
4. Using features to predict GMDS scores
5. Using features to predict MDAT scores
6. Using IRT (item response theory) to generate scores
7. Future Work
8. Understanding tasks and feature extraction

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 ASD/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 (prev slide), 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

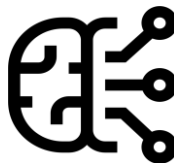
Pipeline



Perform Task
on Tablet



Extract
Features



ML Model



GMDS scores

Training



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

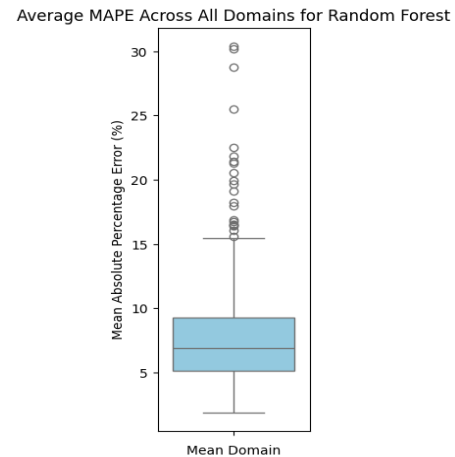
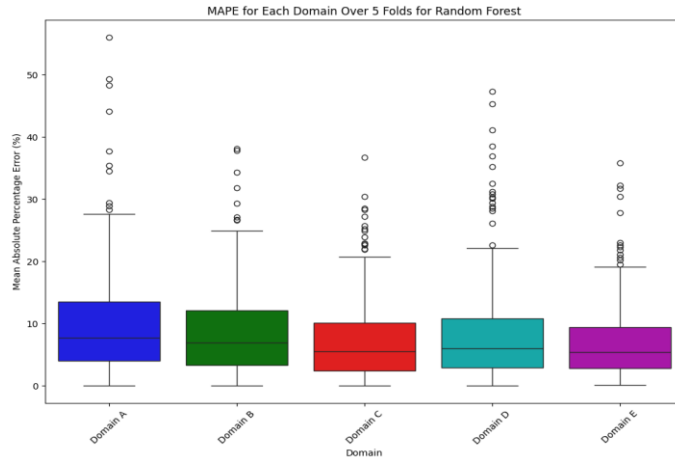
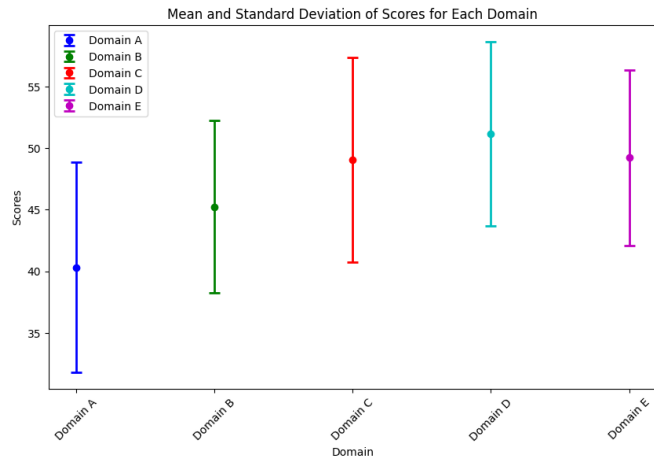
- Linear Regression
- Ridge Regression
- Random Forest
- Gradient Boosting
- AdaBoost
- Decision Tree
- Support Vector Regression
- KNN regressor
- XGBoost

Results (1)

All the results are averaged over the 5 folds:-

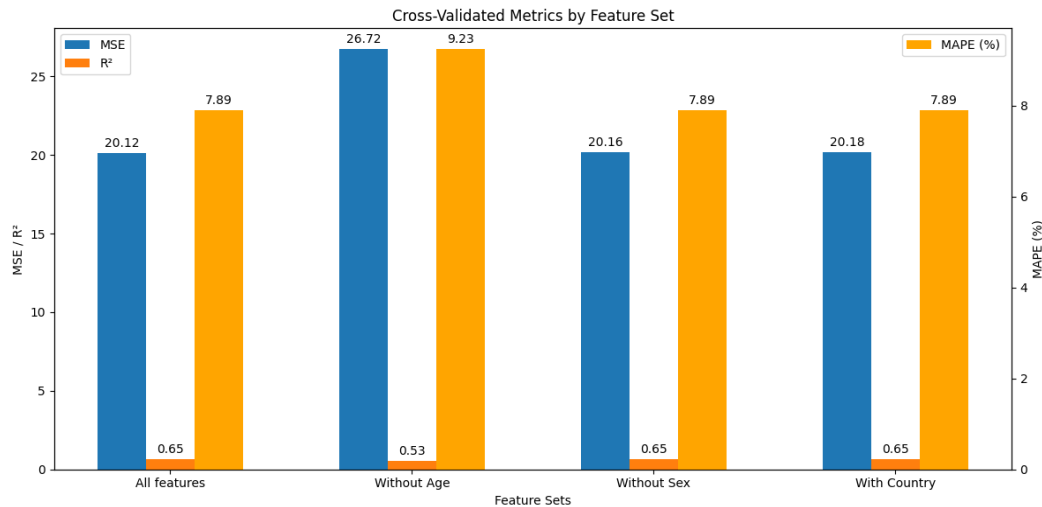
| Model | R2 Score | RMSE | MAPE |
|---------------------------|-------------|--------------|--------------|
| Linear Regression | 0.60 | 22.31 | 8.31% |
| Ridge Regression | 0.62 | 21.56 | 8.19% |
| Random Forest | 0.64 | 20.12 | 7.88% |
| Gradient Boosting | 0.61 | 21.78 | 8.16% |
| AdaBoost | 0.63 | 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

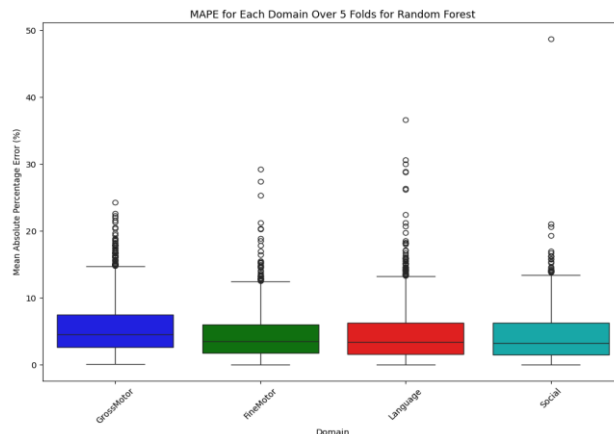
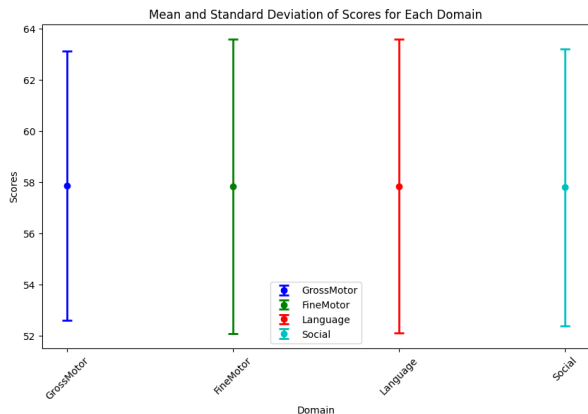
- Linear Regression
- Ridge Regression
- Random Forest
- Gradient Boosting
- AdaBoost
- Decision Tree
- Support Vector Regression
- KNN regressor
- XGBoost
- **Neural Network**

Results (1)

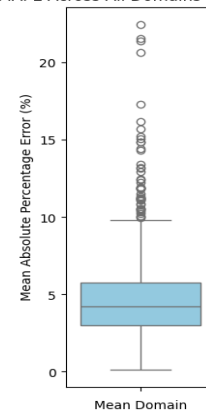
All the results are averaged over the 5 folds:-

| Model | R2 Score | RMSE | 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)



Average MAPE Across All Domains for Random Forest



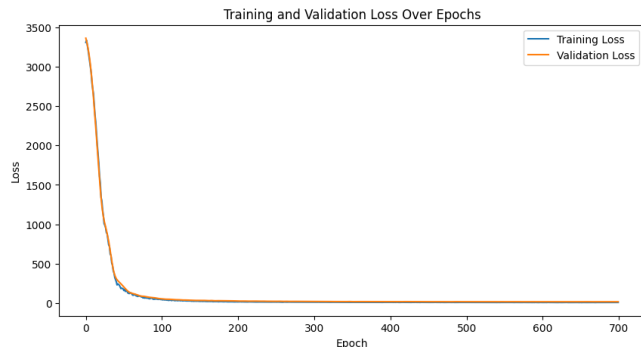
- First plot shows the mean and standard deviation of MDAT scores over the 4 domains
- Second plot shows the box plot for MAPE over the test samples for all the 4 domains
- Third plot shows the box plot for MAPE over the test samples averaged over the 4 domains
- **Model can predict MDAT scores with 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 | 0.05 |

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

- These features should be enough to assess development
- Tasks in tablet are designed to emulate psychometric tests

Item Response Theory (IRT)

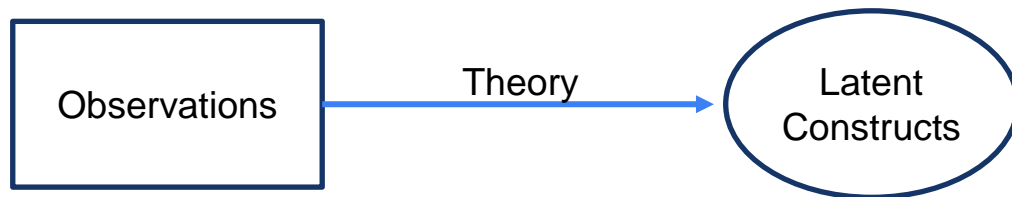
What is IRT?

- 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 |
|---|
| <ul style="list-style-type: none">• Often a test/questionnaire consisting of several 'items'• Could be yes/no questions (could be task responses in our case!) |

| Measurement Theory |
|---|
| <ul style="list-style-type: none">• Participant has an unobserved trait e.g. intelligence, knowledge, anger etc.• Output of measurement tool is mapped to unobserved tool using some 'scaling' |

| Summary |
|--|
| <p>Questionnaires often involve mapping responses onto unobserved traits that are assumed to be continuous</p> <p>But why not just add the responses to get the score</p> <p>Every response is not same... Answering 5 easy questions will results in different intelligence trait than answering 5 difficult question</p> |

Example (Reference)

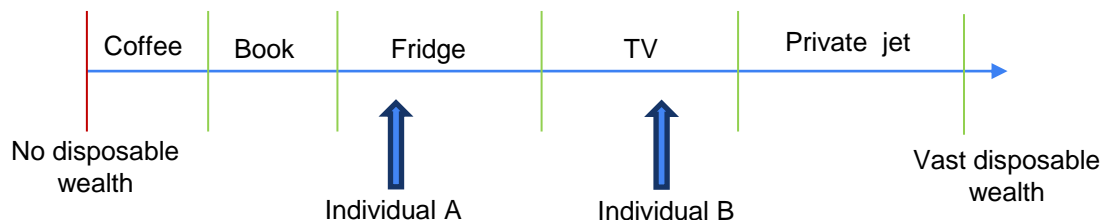
Problem Statement

To measure perceived disposable wealth of person

| Questionnaire |
|--|
| <ul style="list-style-type: none">• 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 |

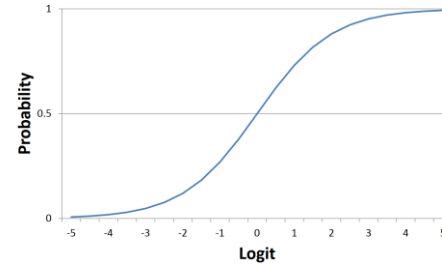
Moving to Probability Space : Rasch/1-Parameter model

$$\text{Logit}_{\text{person,item}} = \text{Wealth}_{\text{person}} - \text{Cost}_{\text{item}}$$

Not between [0,1]

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

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



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

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

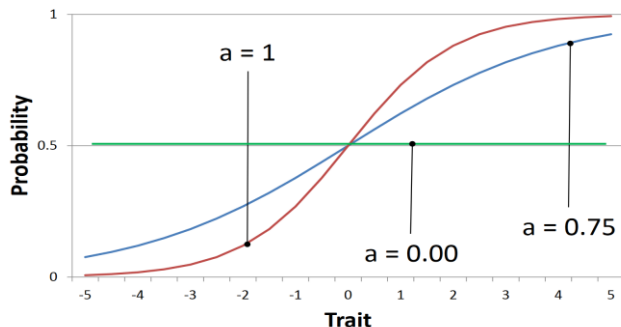
Y_{ij} = Logit of Response by person j for item i ,

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,

$$\begin{aligned} Y_{points\ inside, child\ 1} &= \theta_{child\ 1} - b_{points\ inside} \\ Y_{points\ inside, child\ 2} &= \theta_{child\ 2} - b_{points\ inside} \\ Y_{points\ inside, child\ 3} &= \theta_{child\ 3} - b_{points\ inside} \\ &\dots \end{aligned} \quad \left. \begin{array}{l} \text{Fixed} \\ \text{effect} \end{array} \right\}$$

.....  **Random effect**

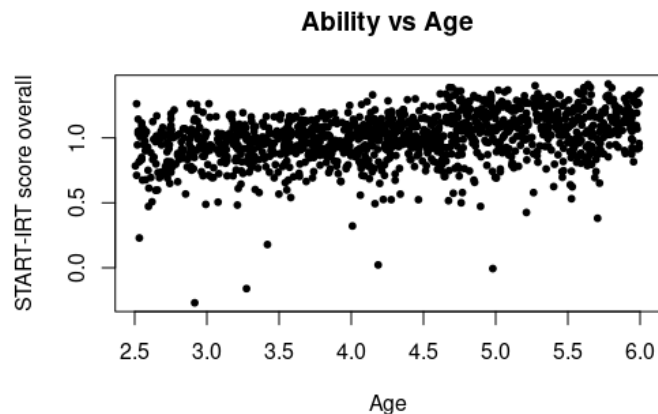
So finally, the final equation would be,

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

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

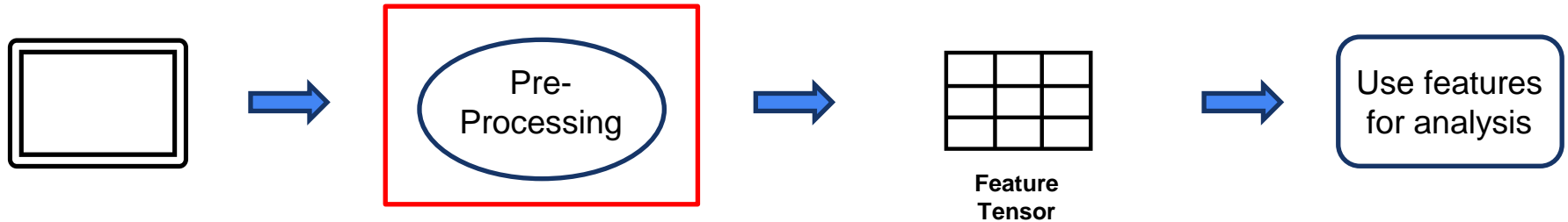
Why is correlation not good?

- IRT assumes monotonicity, 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

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,

Feature description

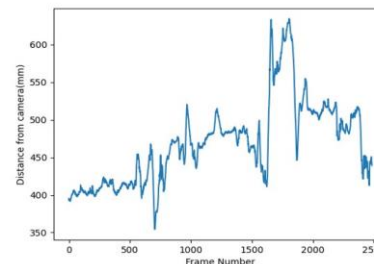
Get the distance of face from the camera using a video, children are instructed to watch the wheel, while their video is recorded on the tablet



Black and white wheel



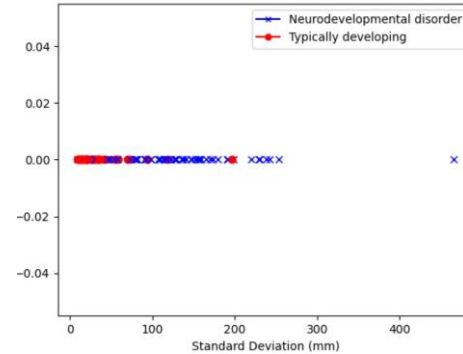
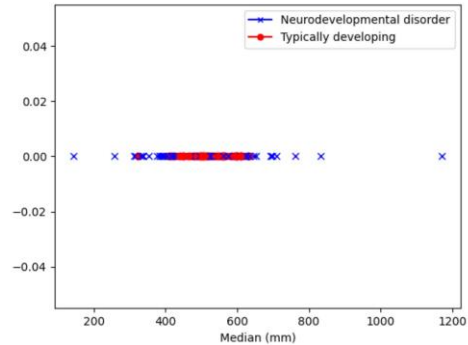
**Video recorded from
front camera**



Distance vs Frame no

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

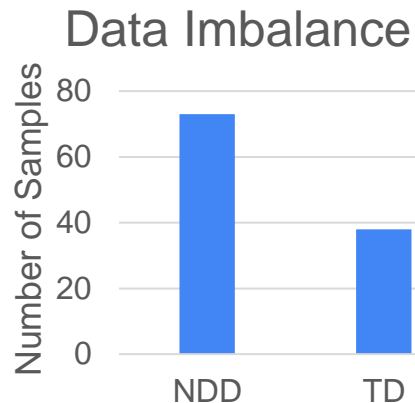
Training Setup

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 |



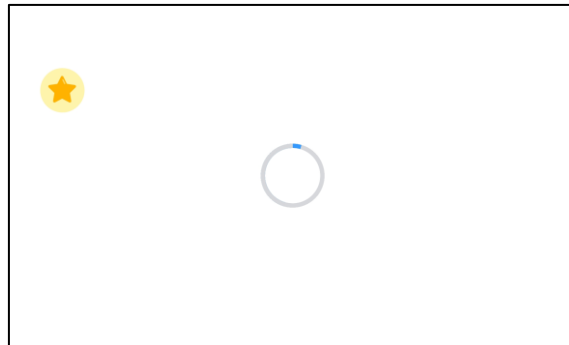
Delayed Gratification Task

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

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

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.

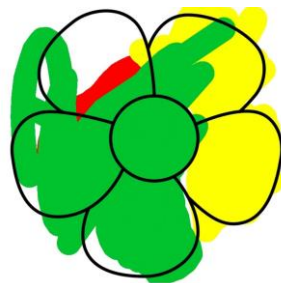
$$\text{Proportion face} = \frac{\text{No of frames with a face}}{\text{Total No of frames}}$$

Coloring Task

What is the task?

Children are presented with simple outline figures and are instructed to carefully fill them with colors

How features are calculated?



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>