Rate My Tweet: Understanding Comparative Judgement in the Wild

Andy Gray
445348

Submitted to Swansea University in partial fulfilment
of the requirements for the Degree of Master of Science

Swansea University
Prifysgol Abertawe

Department of Computer Science
Swansea University

30th September 2021

Declaration

This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ..... (candidate)

Date .....

Statement 1

This work is the result of my own independent study/investigations, except where otherwise stated. Other sources are clearly acknowledged by giving explicit references. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure of this work and the degree examination as a whole.

Signed ..... (candidate)

Date .....

Statement 2

I hereby give my consent for my work, if accepted, to be archived and available for reference use, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

I would like to dedicate this work to . . .

Abstract

In your abstract you should aim to summarise the core contributions of your work in the context of the problem domain. Start by outlining the domain and the problems posed within it. Discuss how the methods you focus on approach the relevant problems. You should end your abstract by concretely stating the tangible outputs and deliverables you have created in order to complete your work on this document, and whether those outputs represent and improvement or alternative approach to existing methods.

Your abstract should be a couple or so paragraphs long, and roughly approximate the order and flow you then use for structuring the main document. If a viewer has read your abstract then they should already understand at a high level what it is you have created and delivered, and whether it is better than or comparable to existing methods. If your project is driven by a research hypothesis then the reader should know what that is at a high level from this section. Reading on, little should surprise the viewer.

For paper submission of your thesis you should physically sign your name and add the date for each of the above declaration statements (black ink preferred). For digital submissions it is normally enough to simply type your name (see custard.cls), though you should sign and date them digitally using a touch or stylus input if at all possible.

Acknowledgements

This is an opportunity to acknowledge and thank those who have supported you throughout your studies. Friends and colleagues who you have studied alongside, your families, and your mentors within the department are the usual suspects.

Contents

Todo list

Chapter 1

Introduction

1.1 Motivations

For the prior eight years, we have had involvement in some form of an educational environment. Seven of these years involve being a teacher within secondary and sixth form schools. While the focus of teaching is perceived to create lessons for students to learn and grow, we found more and more as the years went on that this wasn't the case. The focus was actually on providing reports about the students, which required data about the students from formal assessments. While having assessments to gauge the level that a student is at is an essential part of education. However, creating, marking, analysing and providing feedback for 30 students or more per class is a time-consuming task. Therefore, this assessment practice takes away the educators' time to do what is essential, creating meaningful lessons tailored for the students.

Therefore, our motivation is to create a tool for educators that will empower them to allow technology to do what it is good at and focus on what they are good at, creating and delivering lessons. To shape future generations views.

1.1.1 Objective

1.2 Overview

1.3 Contributions

The main contributions of this work can be seen as follows:

- **A LaTeX thesis template**

  Modify this document by adding additional TeX files for your top level content chapters.

- **A typesetting guide for useful primitive elements**

  Use the building blocks within this template to typeset each part of your document. Aim to use simple and reusable elements to keep your LaTeX code neat and to make your document consistently styled throughout.

- **A review of how to find and cite external resources**

  We review techniques and resources for finding and properly citing resources from the prior academic literature and from online resources.

Chapter 2

Lit Review

Education and the sharing of knowledge is a powerful tool. In fact, in our opinion the most important skill anyone can have. As a famous quote said, "give a man a fish, and he will starve, but teach him to fish, and he won't be hungry anymore". However, it wasn't until 1918 that education, as most people in England and Wales have experienced, started to come into effect [1].

Education over the years was very much about just giving the knowledge to the students from the teacher. It wasn't until 1988, under the Education Reforms Act 1988, that assessments got introduced. The introduction was through the introduction of the national curriculum in England and Wales [2].

As the curriculum got rolled out, statutory assessments got introduced to education between 1991 and 1995. Key Stage 1 first, followed by Key Stages 2 and 3, respectively [3, 4]. Only for the core subjects of English, Mathematics and Science had the assessments first introduced. The first assessments in Key Stage 1 were a range of cross-curricular tasks to be delivered in the classroom, known as standardised assessment tasks - hence the common acronym 'SATs'. However, the complexity of the use of these meant more formal assessments quickly replaced them [3, 4]. The assessments in Key Stages 2 and 3 got developed using more traditional tests.

To allow teachers to judge students' attainment, taking tests became the main assessment form in key stage 3. While assessments were the main form, educators were also able to assess their students with other means against the targets set for attainment within the national curriculum [4]. The teacher and assessment outcomes got used on a scale with key learning milestones expected at different ages. A key stage level indicated the result

for the students progress. The model was used throughout the next few years until 2005 when the role of tests in KS1 got downgraded to just being an internal support tool to teachers, and in then 2008, the government decided to remove tests in KS3 [4].

This model continued, with minor adjustments to reflect the changing content of the National Curriculum, up to 2004. From 2005, the role of the tests got downplayed at Key Stage 1, with tests being used only internally to support teacher assessment judgements [5]. Further changes came in 2008 when the government announced that testing in Key Stage 3 was to get scrapped altogether [6].

However, with a change of government party, the Conservative party taking power from the Labour party brought about new changes to how education's focuses and pedagogy methods would get conducted. In 2014 the system of attainment levels was removed, creating the educational shift of "Assessing without level" [7]. However, within schools, it was being referred to as 'life after levels'. Especially by our educational colleges and us at the time. Which was the follow up to the changes in the national curriculum in 2013 [7]. The changes within the national curriculum brought a greater focus on more traditional style GCSE academic subjects while reducing the focus on perceived technical labour style jobs. The new curriculum direction created more emphasis on the final exam outcomes at the stages of GCSE and A-Level.

2.1 The Purpose of Assessment, Marking and Feedback in Education

As we have established, assessments became a staple of the UK educational system in 1988. While the term assessments are not usually defined, the word 'assess' is typically associated with measuring, determining, evaluating, and judging [8].

While there can be multiple reasons why educators assess students, assessments aim to serve a purpose to both the teacher and the student in the process. These include: giving feedback to teachers and learners; providing motivation and encouragement; to boost the self-esteem of the pupils; a basis for communication; a method to evaluate a lesson/training method/scheme of work/ curriculum; to entertain [8]. Additionally, the assessment also creates other opportunities to rank students; a method to select and filter students, allocate students a particular pathway or educational direction, or as a way to discriminate or choose between students for a given set reason [8].

2.1.1 Traditional Methods of Assessment and Feedback

There are four main categories of assessment. These are diagnostic, formative, summative, and national assessments [8, 4]. However, it is essential to note that national assessments do not get used within everyday aspects of teaching and learning. This term is the name given to the critical exams like SATS, GCSE and ALevel exams taken nationally. Therefore we will focus on the other three main ones.

Diagnostic assessment is what gets referred to as pre-testing [8]. Educators use this technique to get a base level of knowledge of the students they have inherited. This method is good for showing the progress of attainment over time by having an initial base test. Teachers can then show how well the students have progressed over time with their improvements over the term. This base assessment also provides the teacher with crucial information - the current ability of every student's knowledge. Through knowing this current level of knowledge, teachers can adapt the coming lessons and provide suitable differentiation and scaffolding within the lessons to allow each student to succeed as much as possible. However, we also experienced, within our time as an educator, the technique getting used to create baseline narratives. Teachers were using them to show that the student's knowledge wasn't at the expected level when inherited by the teacher at meetings or performance management reviews. Therefore, being used as a counter-act measure tool by the teacher, if they find themselves being accused of letting the students' performance slip, by trying to counter-act by implying the students were not at the required level in the first place.

The second method, formative assessment, is also known as 'assessment for learning (AFL)' [8, 4]. This method has become one of the main tools for a teacher in terms of assessment and feedback. AFL allows the educator to assess the students' understanding of a topic on the fly during a lesson without a summative assessment. As a result, allowing the teacher to spend more or less time if the students do or don't get the topic, even if they planned more or less time for that topic. Therefore, ensuring that the teaching is not getting carried out for teaching sake. Thus, the emphasis is less on measurements and more on actual learning. AFL can involve using several techniques: teacher assessment - through in-class questions, marking books; to the students assessing their work called self-assessment, or peer assessment - where the students evaluate each other's work [8].

AFL has many values for teachers and students. Within Black and William's paper. 'Inside the black box [9]' discovered that AFL provides massive learning gains, especially

with the low attainer groups. Black and William found that AFL and the use of peer assessment raised motivation and self-esteem across the board, but even more so in the low attainers. With the addition of peer assessment being extra valuable to the students. This form of feedback is effective as the feedback will most likely be given back to the students in a manner that they are more familiar with, informs of language and wording. Therefore in a way that makes more sense to them and having the most impact on their learning [10, 9].

The third method is a summative assessment, also known as 'assessment of learning (AOL) [8]. This type of assessment happens at the end of a teaching unit or topic. It gets used to gain insights into what the students have learnt within the subject covered or the course. Its purpose is to give a student a mark, grade or ranking. Usually, this is the grade that is mainly focused on, as it is the metric that will impact the school the most in terms of league performance tables regarding GCSE and A-level results. From our experience, summative assessments are carried out regularly within schools. This assessment method tends to get used to getting a snapshot of the students of whit if a moment like, if they were to take the test now, what would they get? By seeing the results, educators can see if students need to attend intervention or if they are performing as expected or even better. With so much riding on these results, for schools and teachers performance management reviews, a lot of emphasis on put into trying to predict the final results for students. We have seen it put a lot of pressure on the teachers and the students and ultimately creates a very stressful environment, which is not the best environment for learning.

2.1.2 Why Traditional Traditional Marking and Feedback Methods are Effective

2.1.3 The Negative Aspects of Traditional Marking and Feedback Methods

2.2 Comparative Judgement

2.2.1 What is Comparative Judgement

Comparative judgement is a mathematical way to determine which observation item is better than the other item also being observed compared to each other. This method was first proposed in 1927 by Louis Leon Thurstone, a psychologist, under the term "the law of comparative judgement" [11, 12]. In modern-day terminology, it gets more aptly described as a model used to obtain measurements from any pairwise comparison process.

Examples of such methods are comparing the perceived intensity of physical stimuli, such as the weights of objects, and comparing the extremity of an attitude expressed within statements, such as statements about capital punishment. The measurements represent how we perceive things rather than being measurements of actual physical properties. This kind of measurement is the focus of psychometrics and psychophysics. <wikipedia>

In more technical terms, the law of comparative judgment is a mathematical representation of a discriminal process. This process involves a comparison between pairs of a collection of entities concerning multiple magnitudes of attributes. The model's theoretical basis is closely related to item response theory and the theory underlying the Rasch model. These methods are used in psychology and education to analyse data from questionnaires and tests. <wikipedia>

While comparative judgement is a technique that has been around for almost 100 years, it wasn't until the early nineties that this technique got proposed for use within an educational setting. This first proposal was by Politt and Murry [13], who conducted a study where they tested candidates on their English proficiency within Cambridge's CPE speaking exam. The judges watched 2-minute videos and judged which one out of a pair of videos they deemed better at the requested task in the exam. However, before this, in the ninety seventies and eighties, comparative judgement was presented as a more theoretical basis for educational assessments [14].

With the momentum of his findings, Politt then presented comparative judgement as a tool for exam boards to use to be able to compare the standards of A-Levels from the different exam boards, replacing the direct judgement of a script that was at the time currently being used [15]. In his papers titled, "Let's Stop Marking Exams" [16], he presents a valid argument for using comparative judgement, with the advantages it brings over some traditional types of marking.

Politt, in 2010, also presented a paper at the Association for Educational Assessment – Europe. It was about How to Assess Writing Reliably and Validly. Politt presented evidence of the extraordinarily high reliability achieved with Comparative Judgement in assessing primary school pupils' skill in first-language English writing [17].

2.2.2 The Logic Behind Comparative Judgement and What it Aims to Do

How comparative judgement works is to present two options to a marker. The marker then gets asked to pick which one of the two options they think is the better one. The

marker will get presented with all possible combinations available, each time picking which one they think is the better one out of the two. An outputted score is then presented based on the method used. The original method, the Law of Comparative Judgement (LCJ), follows the formula:

$$S_i - S_j = x_{ij}\sqrt{\sigma_i^2 + \sigma_j^2 - 2r_{ij}\sigma_i\sigma_j},$$

Figure 2.1

S_i is the psychological scale value of stimuli i

However, an alternative version derived from Louis Leon Thurstone, referred to as the "Pairwise Comparison" [12], will provide an output based on the difference between the quality values is equal to the log of the odds in respect to object-A will be object-B. This formula gets represented as: $\log \text{odds}(A \text{ beats } B \mid v_a, v_b) = v_a - v_b$.

$\Pr\{X_{ji} = 1\} = \frac{e^{\delta_j - \delta_i}}{1+e^{\delta_j - \delta_i}} = \sigma(\delta_j - \delta_i)$

.

2.2.3 How effective is Comparative Judgement at Providing Feedback?

2.3 Related Work

2.3.1 Subsection all similar work

2.3.2 Comparison of similar work

Chapter 3

Methodology

3.1 Tools

To create the web application and insights from the tweets, we required to use several tools. It is a requirement that we develop a full-stack web application with a user UI, an area to input the user's judgements on the tweet, store the results using a database, and extract information from the tweets using NLP techniques. Several factors within the final application needed to be satisfied for the tools to be appropriate for use.

3.1.1 Programming Language

While many programming languages can handle creating a full-stack application and conducting ML, for example, Java, Php and JavaScript. We decided to use the Python language. We decided upon Python due to our familiarity with it over the other main languages and its versatility. We made this decision because Python can make full-stack applications with the use of additional libraries, as well as handle most NLP ML tasks using libraries like NLTK, SpaCy, Sci-Kit Learn and TensorFlow.

3.1.2 Libraries

3.1.2.1 Web Application

For creating the web application, there were two main libraries available. These were Django and Flask.

Django is a high-level Python Web framework that encourages rapid development and clean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of Web development, so you can focus on writing your app without needing to reinvent the wheel. It's free and open source [18].

While Flask is a small framework by most standards—small enough to be called a "micro- framework," and small enough that once you become familiar with it, you will likely be able to read and understand all of its source code [19].

Flask has three main dependencies. The routing, debugging, and Web Server Gateway Interface (WSGI) subsystems come from Werkzeug; the template support is provided by Jinja2; and the command-line integration comes from Click. These dependencies are all authored by Armin Ronacher, the author of Flask [19].

Flask has no native support for accessing databases, validating web forms, authenticating users, or other high-level tasks. These and many other key services most web applications need are available through extensions that integrate with the core packages. As a developer, you have the power to cherry-pick the extensions that work best for your project, or even write your own if you feel inclined to. This is in contrast with a larger framework, where most choices have been made for you and are hard or sometimes impossible to change [19].

After experimenting with the two frameworks, we decided upon Flask. Flask got decided upon because of the short time frame to put the project together. Additionally, the lightweight nature of the framework also played a fact as we believe that as this will be just an initial prototype, all the other requirements that Django requires would be unessential additionals to the project. Therefore, taking focus away from what we believe is the main focus.

3.1.2.2 NLP Tasks

3.1.3 IDE

3.2 Software Development Life Cycle Methodology

3.3 Data Set

3.3.1 Data Capture Method

3.3.2 Pre-Processing

Chapter 4

Results and Discussion

In this document we have demonstrated the use of a LaTeX thesis template which can produce a professional looking academic document.

4.1 Contributions

The main contributions of this work can be summarised as follows:

- **A LaTeX thesis template**

  Modify this document by adding additional top level content chapters. These descriptions should take a more retrospective tone as you include summary of performance or viability.

- **A typesetting guide for useful primitive elements**

  Use the building blocks within this template to typeset each part of your document. Aim to use simple and reusable elements to keep your document neat and consistently styled throughout.

- **A review of how to find and cite external resources**

  We review techniques and resources for finding and properly citing resources from the prior academic literature and from online resources.

4.2 Future Work

Future editions of this template may include additional references to Futurama.

1: Add this yourself and submit a pull request?

Chapter 5

Conclusions and Future Work

In this document we have demonstrated the use of a LaTeX thesis template which can produce a professional looking academic document.

5.1 Contributions

The main contributions of this work can be summarised as follows:

- **A LaTeX thesis template**

  Modify this document by adding additional top level content chapters. These descriptions should take a more retrospective tone as you include summary of performance or viability.

- **A typesetting guide for useful primitive elements**

  Use the building blocks within this template to typeset each part of your document. Aim to use simple and reusable elements to keep your document neat and consistently styled throughout.

- **A review of how to find and cite external resources**

  We review techniques and resources for finding and properly citing resources from the prior academic literature and from online resources.

5.2 Future Work

Future editions of this template may include additional references to Futurama.

2: Add this yourself and submit a pull request?

Bibliography

[1] UK Public General Acts, "Education act 1918," 1918.

[2] ——, "Education act 1988," 1988.

[3] D. Hutchison and I. Schagen, *How reliable is National Curriculum assessment?* NFER, 1994.

[4] J. Dillon and M. Maguire, *Becoming a teacher: Issues in secondary education*. McGraw-Hill Education (UK), 2011.

[5] BBC News. (2004) Primary school tests toned down. [Online]. Available: http://news.bbc.co.uk/1/hi/education/3656244.stm

[6] ——. (2008) Tests scrapped for 14-year-olds. [Online]. Available: http://news.bbc.co.uk/1/hi/education/7669254.stm

[7] Department for Education. (2013) Assessing without levels. [Online]. Available: https://webarchive.nationalarchives.gov.uk/ukgwa/20130802141012/https://www.education.gov.uk/schools/teachingandlearning/curriculum/nationalcurriculum2014/a00225864/assessing-without-levels

[8] J. Wellington, *Secondary education: The key concepts*. Routledge, 2007.

[9] P. Black and D. William, "Inside the black box: Raising standards through classroom assessment. phi delta kappam," 1998.

[10] H. Torrance and J. Pryor, *Investigating formative assessment: Teaching, learning and assessment in the classroom*. McGraw-Hill Education (UK), 1998.

[11] L. L. Thurstone, "Psychophysical analysis," *The American journal of psychology*, vol. 38, no. 3, pp. 368–389, 1927.

[12] ——, "A law of comparative judgment." *Psychological review*, vol. 34, no. 4, p. 273, 1927.

[13] A. Pollitt and N. L. Murray, "What raters really pay attention to," *Studies in language testing*, vol. 3, pp. 74–91, 1996.

[14] D. Andrich, "A rating formulation for ordered response categories," *Psychometrika*, vol. 43, no. 4, pp. 561–573, 1978.

[15] P. Newton, J.-A. Baird, H. P. Harvey Goldstein, and P. Tymms, "Paired comparison methods," 2007.

[16] A. Pollitt, "Let's stop marking exams," 01 2004.

[17] ——, "Abolishing marksism and rescuing validity," *International Association for Educational Assessment, Brisbane, Australia. http://www. iaea. info/documents/paper_4d527d4e. pdf*, 2009.

[18] Django. (2021) Meet django. [Online]. Available: https://www.djangoproject.com/

[19] M. Grinberg, *Flask web development: developing web applications with python*. " O'Reilly Media, Inc.", 2018.

Appendix A

Implementation of a Relevant Algorithm

```
#include <stdio.h>

int main(int argc, char *argv[]) {
  printf("Hello world.\n");
  return 0;
}
```

Listing A.1: An implementation of an important algorithm from our work.

Appendix B

Supplementary Data

The results of large ablative studies can often take up a lot of space, even with neat visualisation and formatting. Consider putting full results in an appendix chapter and showing excerpts of interesting results in your chapters with detailed analysis. You can use labels and references to refer the reader here for the full data.