Interaction Technology

HIGH SCHOOL STUDENTS INTERACTION WITH A CHEMICAL LEARNING APP

d.g.b. van Rijswick

MASTER'S ASSIGNMENT

Committee:
Robby van Delden
Maurice van Keulen

July, 2024

Interaction Technology

University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Abstract

This research aimed to develop and evaluate an application designed to assist high school students in learning to write chemical formulas by hand. The study was structured into three stages: input methods, Optical Character Recognition (OCR) processing techniques, and feedback mechanisms. The main research question addressed was: "How can an application be designed to assist high school students in learning to write chemical formulas by hand effectively integrate input methods, optical character recognition (OCR) processing, and feedback mechanisms?"

In the input stage, three methods were tested: writing on plain paper, writing in grid boxes, and writing directly on a tablet. The results were analyzed taking into account the demographics, indicating significant differences between schools in terms of the preferred method. These differences highlight the potential benefits of tailoring educational practices to suit the specific needs and contexts of different school environments to enhance student engagement and learning experiences.

The OCR processing stage involved evaluating various models to recognize handwritten chemical formulas. The model using the MNIST and A-Z datasets achieved a high accuracy of 97%, but it required additional modifications to handle chemical formulas effectively. Models specifically trained on chemical formulas faced challenges due to insufficient data, underscoring the importance of having larger and more diverse datasets for accurate OCR performance. Future research should focus on expanding these datasets and exploring advanced machine learning techniques to improve OCR model accuracy for educational applications.

The feedback stage assessed three methods: providing the correct answer, marking mistakes, and offering detailed explanations. In giving the correct answer, participants reflected on their answers after being shown the correct answer, which prompted self-assessment and critical thinking. Common issues identified included problems with subscript and superscript notation, capitalization errors, chemical formula inaccuracies, and various other comments. The Technology Acceptance Model (TAM) and the VARK learning styles test provided comprehensive data on student preferences and learning styles, revealing a preference for structured and incremental feedback.

Despite several limitations, including app malfunctions and insufficient training data for OCR models, the study provides valuable insights into the development of educational tools for chemistry. One notable limitation was the incomplete collection of ethics consent forms, which restricted the participant pool. Future research could address this by digitizing the consent form process to streamline submissions. Additionally, ensuring that experiments can be conducted efficiently with whole classes or in staggered groups is crucial for maximizing data collection within limited time frames.

The findings highlight the crucial role of user-friendly input methods, specialized OCR datasets, and progressively detailed feedback in enhancing students' learning experiences. These insights will guide future research and the development of more effective educational applications.

Contents

1 Introduction

In the ever-evolving field of education, the integration of technology has become a major focus of innovation. As we stand at the crossroads of traditional teaching methodologies and the digital age, it becomes important to explore how technologies can enhance the learning experience. In the classrooms of today, many programs are being used to aid the management of the class, teach and test knowledge, and as tools to create for the class. However, there are few technologies specifically designed to teach skills to students, particularly in the area of handwritten chemical formulas, according to Zydney and Warner (2015). This project aims to bridge this gap by delving into the realm of interaction technology and education, focusing on the development of an app designed to aid high school students in mastering the intricacies of writing chemical formulas by hand.

While digital tools are increasingly prevalent in standard education Zydney and Warner (2015), significant areas for improvement remain. One of these areas is chemistry education, especially writing chemical formulas by hand, since a small mistake here can completely change the meaning. High school chemistry teachers that joined our focus group emphasized the importance of this skill, noting its crucial role in the field and its necessity for the national final exam. These handwritten formulas can be accurately detected using Optical Character Recognition (OCR) if the system is properly adapted to handle the nuances of chemical notation. This project will initiate the development of an app that assists high school students in learning how to correctly write chemical formulas by hand, aiming to make it easier for students to master this essential skill.

We begin by exploring the research question: "What do high school students prefer when interacting with an app designed to improve their ability to write chemical formulas by hand?" This multifaceted research will focus on the means of input and feedback provided by the app. These objectives will be addressed through a meticulous exploration of various input methods, ranging from traditional paper to digital tablets, and by testing diverse feedback approaches, including self-explanation prompts, marking mechanisms, and sentence-based feedback.

Additionally, this research will address the technical challenge of "How can an OCR system be created for chemical formulas?" This will involve developing a model based on the input data from the app, aiming to achieve an accurate representation of handwritten chemical formulas.

The main research question this report seeks to answer is: "How can an application be designed to assist high school students in learning to write chemical formulas by hand effectively integrate input methods, optical character recognition (OCR) processing, and feedback mechanisms?" This question will be examined through three sub-research questions. The first is: "Which input method is preferred by high school students?" The second is: "What is the accuracy and effectiveness of different OCR models in recognizing and processing handwritten chemical formulas?" The third is: "What is the influence of different feedback methods on the perception of the students?"

The background section delves into existing studies and technologies, setting the stage for the current research. Following this, an overview of a focus group conducted with high school teachers provides valuable perspectives into the field of science education. The Methodology section outlines the detailed plan of action, from participant selection to data analysis. The Results section presents the findings of the research, highlighting key trends and observations. Finally, the Conclusion section draws conclusions from the results, and the Discussion section explores the implications for future research and educational practices.

As we embark on this exploration, the goal is not just to design an app but to create an educational experience that meets the needs and preferences of high school students learning the details of handwritten chemical formulas. Through this research, we aim to contribute valuable insights to the broader field of technology-mediated learning, specifically focusing on the use of OCR for chemical formulas.

2 Background

In this section, we will present the research conducted in the fields of technology in education, optical character recognition (OCR), and feedback for students. The papers discussed in this section were identified using search terms such as technology integration in education, mobile learning, handwritten formula detection using OCR, formula detection using OCR, feedback in education, and effective feedback. The search results were first scanned based on their titles and abstracts. Those deemed most relevant were then read in detail, and the most prominent studies are summarized in this chapter.

2.1 Technology in Education

An important goal of this research is to develop a system that helps students learn to write chemical formulas by hand accurately. The initial step is to determine the most effective way to implement technology in science education and to understand the key factors that contribute to its success.

It is first important to look at essential factors for including technology into the curriculum. Hamidi and Chavoshi (2018), Masrek and Samadi (2017) and Sarrab et al. (2016) researched this by investigating what students think about this subject. They found that there were a few key factors that would contribute to a successful integration into the lessons. The first factor is *ease of use*, as new technologies should not require a lot of time and effort to be used effectively. The second factor is *trust*; if students and teachers do not trust the technology, they are unlikely to use it. The investigation into these two factors revealed that the context in which the technology is used significantly impacts its ease of use and overall effectiveness. Furthermore, ease of use has a big influence on the perceived usefulness. They also found that behavioral intention factor, meaning the acceptance of a technology among students, is not related to the individual student, ease of use or usefulness of the technology, however the trust in the technology has a big influence on the behavioral intention factor.

Wohlfart et al. (2023) and Badia et al. (2014) looked at how teachers think about using technology in the classrooms. This paper found that teachers are highly interested in incorporating technology into their lessons and recognize the value it can bring. However, they also mentioned several challenges to implementation, such as time constraints and high workloads. Additionally, there is a problem with the infrastructure of the education system that supports the technologies teachers want to use. This includes issues with digital platforms, apps for communication, and digital teaching materials. Furthermore, there is a lack of technological support to effectively integrate these tools. Che Rose and Lawrence (2008) discussed that it is important for policy makers and school administrators to create an environment where teachers can easily use digital tools for education.

Now that the perspectives of students and teachers are known it is good to look at what technologies already exist and can be used. Zydney and Warner (2015) looks into mobile apps specifically for science learning, while Riojas et al. (2012) also looks at physical technologies. Zydney and Warner (2015) categorized the apps into four categories: the first category is place-based data collection tools, which use GPS to facilitate learning at specific locations; the second category is games and simulations, which involve learning in a virtual environment; the third category is learning management systems, which help manage lessons, activities, and communications; and the fourth category is productivity tools, which are used for creating educational content, such as digital presentations, documents, or interactive projects.Riojas et al. (2012) divided technologies based on deductive or inductive learning methods used in the technologies. Deductive learning (e.g., learning by applying general rules to specific examples) and inductive learning (e.g., learning by observing specific examples and then forming general rules) were both considered. These papers also noticed that these systems focus a lot around sharing information and that these apps focus mainly on teaching knowledge and testing it instead of teaching skills. One problem the paper found was that the apps reference teaching theories, but they do not actually use this information in the app. They concluded that future developers should include a social aspect, implement learning theories, and find innovative technologies to incorporate into a useful system. This recommendation stems from the observation that current apps primarily focus on sharing information rather than fostering collaborative learning and interaction.

Continuing on the subject of collaboration, Fu and Hwang (2018) and Vali (2023) looked into how the standard collaborative learning methods can be done using technologies by doing a systematic review of different journal papers. The paper finds that the portability of mobile technologies is a significant advantage for enabling collaboration in various settings, such as within a classroom, between different classrooms, throughout an entire school, and even outside the school environment. They notice that this is very useful to facilitate collaborative learning and that it can help students to gain knowledge, skills and beliefs. They conclude that using social interaction improves the learning abilities of the students.

These existing technologies use different forms of interaction and Alin et al. (2012) has looked into the different forms of interaction with technologies by systematically analysing different technologies. This paper discusses how people interact with technologies and the key features of different technologies. It shows the importance of the mobility of phones and the many features that can be integrated into mobile applications. One of the features is the touchscreen which can be used to write directly on the phone or tablet. Mohammed and Karagozlu (2021) looked at different forms of interaction interfaces by reviewing the design directions of multiple papers. They found that the static interfaces which were based on desktop paradigms did not succeed, because modern systems are more focused around mobile phones.

To conclude the opinions and expectations of the teachers and students is positive although they have some reservations on the implementations, so it is important that an app can be trusted and is easy to use.One of the important aspects of learning is collaboration, and a key feature of modern technologies is the ability to stay connected wherever you go. Therefore, it looks very promising to create an app for an educational dedicated tablet, phone or laptop that can leverage these features to help students learn.

2.2 Optical Character Recognition

One important aspect of this project is to transform the handwritten chemical formulas into digital representations. This can be done using Optical Character Recognition (OCR), which analyses images of the handwritten text and converts it into a digital representation.

First it is important that we know what OCR is. Memon et al. (2019) has done a systematic review of papers on OCR in the IEEE database and Mohammed and Karagozlu (2021) describes the basics of an OCR system. Mohammed and Karagozlu (2021) defines an OCR system to have three stages, the first stage is to scan the image, the second stage is to decode the image into text, and the third stage is the output interface. The first stage is simply having a well lit document scanned using a lens with a detector such as a camera. The second stage contains document analyses, followed by the character recognition in conjunction with a contextual processor to test the result of the character recognition and provide feedback. The third stage is simply showing the recognition results. Memon et al. (2019) found that there are many methods or decoders to do OCR such as Artificial Neural Networks, Kernal methods, statistical methods, template matching, and structural pattern recognition. Which all aim to do the same thing, transform an image into a digital text. All of the OCR systems are trained on datasets, mostly on standard datasets, none of which focus specifically on chemistry. As Memon et al. (2019) mentions, eight standard datasets are commonly used in many of the papers. The lack of chemistry-specific datasets will impact the ability of OCR systems to recognize chemical formulas accurately, especially when dealing with the use of superscript, subscript, and the distinction between capital and lowercase letters. They also mentioned the six languages that are used most for OCR training.

An additional part of an OCR system can be to pre-process the data, which Jaiswal et al. (2023) and Yousif (2024) focuses on. Yousif (2024) looked into general pre-processing of data such as normalization, random rotation, and cropping of the image using an automated system, in this case a deep learning model. They observed improved results for accuracy and computation time of the entire OCR system. Jaiswal et al. (2023) focuses on creating a system to first separate the full image into separate images containing only the words with a focus on low-quality data. Using this technique the paper finds that they can increase the recognition from around 55% to 92%, which is a significant improvement when looking at low-quality data. A similar method can be used to identify the individual formulas in chemical equations.

After the pre-processing is done the formulas need to be processed using OCR models. Garst et al. (2023) looked into creating OCR models with custom vocabulary. The paper created a method to train a model on limited data to create a model that can work in a domain specific way. For this purpose they also created a modified decoder to create an estimate of the word or expression based on the custom vocabulary. Orji et al. (2023) also investigated custom vocabulary OCR systems, specifically for image-to-LaTeX conversion, focusing on recognizing mathematical expressions, formatting styles like italics and bold, and distinguishing between capital and lowercase letters. The paper emphasize the role of context aware models and also focused on active learning, meaning that the model will be updated when more information becomes available. The paper also included syntax constraints to make it clearer. These methods could be implemented for chemical equations, it is important to note that a custom OCR model should maintain the mistakes written by the students.

When the handwritten answers have been analysed by OCR it is also possible to do some post-processing like what is done by Hemmer et al. (2023) and Karthikeyan et al. (2021). Hemmer et al. (2023) focused on analysing denoising complexity on numerical texts, by creating an estimator that looks at an optimal denoising method for numerical texts with textual noise and compare this with a normal more complex denoising method. This means that they created deliberate mistakes in the results after the OCR to check how different methods work in correcting these mistakes. Based on this method it might be possible to create a denoising method for chemical formulas. Karthikeyan et al. (2021) looked at using deep learning models for post processing and found a significant improvement regarding word error rate and character error rates.

The results still needs to be analyzed and Naiman et al. (2023) focuses on how to deal with historical documents and the ambiguity that exists within the ground truth, while Lopresti (2009) looks at common errors and the consequences when the results will be used for further analysis. A big problem with historical documents is that the ground truth is not always consequent in how to represent certain characters or combinations, while this does not matter for the meaning it makes the OCR training more difficult. And like the teachers said during the focus group, grading is a gray area and can differ greatly between situations. An example the teachers mentioned was that some answers are marked correct early in the education, while the answer would be wrong later in education. This also leads to moments where the same answer could be marked correct for one student and incorrect for another student in the same class, because the marking is not very rigid especially around edge cases. The errors that appear as a result of problems in the training or execution stage of OCR, such as the ambiguity problem, have an impact on not only the results but also on the following analysis of these results. This impact has been researched by Lopresti (2009) on a large dataset of documents. They created a setup that included an OCR stage, a sentence boundary detection stage, a tokenization stage, and a part-of-speech (POS) tagging stage, which involves grammatical classification to identify the parts of speech for each token. The errors that had the most significant consequences were centered around punctuation and spaces, while ambiguity in the classification of characters primarily affected the POS tagging stage. However, in the context of this project this stage represents the most important part of the analyses.

OCR's primary function is to digitize text, making it an essential tool for converting handwritten or printed material into digital form. Rijhwani et al. (2023) looked into using OCR in transcribing text in this case for the language Kwak'wala and Robert et al. (2024) looked into using OCR to transcribe a natural history collection. Rijhwani et al. (2023) shows how people with and without knowledge of the language transcribe the text. The results are that it makes the process a lot faster, however not everyone preferred using OCR over the normal procedure. For the chemical equations this can be used when the OCR shows a completely different answer then that was written down. This information can then also be used to improve the model using the active learning methods of Orji et al. (2023) to keep improving the model while it is being used. Robert et al. (2024) also looked at creating an assistive tool, but focused their research more on the results of incorporating humans into the process. They found that using humans can indeed be useful, however they also noticed that the accuracy could drop because of human participation. There is a benefit, but the workflow needs to be simple and clear to minimize the chances of errors.

To conclude OCR has many steps and considerations, but can definitely used to digitize the handwritten chemical formulas to analyze the answers. From the pre-processing to making a custom vocabulary that can update when more information is available.

2.3 Feedback

After the OCR has done its work to create a good representation of the written answers, the answers need to be compared to the correct answer and feedback needs to be created. Some forms of feedback that are currently used include scribbles, short comments, or abbreviations. In this section, we will look at research on how students think about feedback, what are effective methods of feedback, and how feedback works for students that have trouble with learning.

Looking at the research from Rowe and Wood (2008) on student perceptions and preferences shows that students can be categorized into two groups, first students that dive deep into the knowledge and use feedback to improve and second students that only care that they pass and ignore feedback when they have passed. This also translates into the form of feedback that they prefer, the first group likes to receive detailed feedback about the content, while the second group prefers positive feedback and the correct answer directly. The research also revealed a difference in feedback satisfaction based on the year of study. Students in earlier years were generally more satisfied with the amount of feedback they received, whereas students in later years expressed dissatisfaction not only with the amount of feedback but also with the type of feedback provided. For younger students, feedback tends to be more frequent and straightforward, while older students prefer more detailed feedback to understand their mistakes better, although they also indicated a preference for more concise feedback due to their busier schedules. Another factor observed was gender, with female students generally expressing a stronger preference for receiving feedback, particularly detailed feedback. This preference might correlate with the trend seen in the earlier years of study, where students showed a greater desire for frequent and straightforward feedback. It could suggest that more female students fall into this group. There was also a general trend found, most students found there was not enough feedback and often felt they received it too late. Vaessen (2021) also looked at student perception of feedback and also found a relation between study methods, study willingness, and how students perceive feedback. With students less interested in studying also perceiving feedback more negatively than students that put more effort in studying.

Nurjanah (2021) confirms this problem that feedback is often not enough, however they also found that feedback generally is perceived as positive. This research found that one of the causes is that the feedback is not clear and does not give more understanding of the materials. They also gave a solution on how to structure feedback, they say that feedback needs to contain three parts, first the goal of the answer, second how is the student doing to get to the answer and finally what does the student need to do. In this way the student knows what is the material, what am I doing correct and what do I need to do to improve.

This method of how to give effective feedback is supported in "How to give effective feedback to your students." by Brookhart (2008) which talks about what feedback is, which parts are important for feedback and how to give this feedback. This book is from the association for supervision and curriculum development in the USA. It mentions common mistakes such as giving feedback without any material content such as that is wrong. This book also talks about the effective parts of feedback, namely that feedback should motivate students to use the feedback and to improve through self reflection as well as by using the feedback of the teachers. The book mentions the four most important aspects of feedback to be timing, amount, mode and audience. The book also gives a more detailed explanation about how to use oral and written feedback. For written feedback the book shows that it is important to know where to write feedback (either close to the evidence, on a standard rubric or using both), the tone of the feedback (give not only negative feedback but also positive and use easy to understand language) and the specificity of the feedback (clear statements that show the mistakes, but leave room for the student to figure out how to improve it).

The research from Bahati et al. (2016) focused on formative feedback (feedback to improve, in contrast to summative feedback which only mentions how well or poorly they did, such as in exams) and noticed that the teacher was in most cases the sole person responsible for providing the feedback. They found that from the teachers perspective oral feedback was more important, however students preferred written feedback since the oral feedback would cost time during lessons. Similar to findings from previous research, students found the feedback to be unclear or insufficient. This was partly because the feedback was not structured, as mentioned by Brookhart (2008), but instead relied on using marks, which refers to simple annotations or grades without detailed explanations.

Pitt et al. (2019) looked into feedback related to students that have trouble with studies to understand what needs to be improved so all students can have a better experience studying. They found that structured feedback incorporating positive aspects works better than negative low content feedback. They also found that it is important to have a close relationship with other students and the teacher to better understand and use the feedback provided. In this way students can discuss the feedback and better learn what the feedback means and how it can be used to improve.

The research in Irwin et al. (2011) focused on the problem that students are not engaged enough with the feedback and the teachers. They found that the structure of the feedback is very important to the engagement of the students. Another important factor that they found is that a grade combined with feedback is a lot less effective then just providing feedback. So they suggest to use technology to first give feedback and let the student understand the feedback before giving them the grade, so that they need to spend time on the feedback and thus get more engagement with the feedback which results in a better understanding.

To conclude, the main problems with feedback are related to how it is given, when it is given, and how students use it. Research shows that it is important for students to interact with feedback to better understand the material. However, it is most crucial for students to understand the feedback so they can engage with it effectively. Additionally, USA-based students have shown a preference for written feedback over oral feedback, which further emphasizes the need for clear and detailed written feedback to facilitate better student engagement and understanding.

3 Expert's Insights

We are focused on improving the way that high school students learn how to write chemical formulas and equations. In order to achieve this goal we believe that it is important that the teachers who teach these classes are involved in the process**?**. This starts with a focus group to create a clear understanding of the current situation, what the teachers expect to be an improvement, and the reasoning of why certain points are important.

3.1 Semi-Structured Focus Group

The focus group is structured to be in a group session with multiple science teachers present to spark discussion and find the core of the current situation, the expected situations and reasoning behind the ideas. To get the answers to these questions the focus group is organized in order of importance. Below are the questions, where the numbered questions are the main target questions. Underneath these questions, some additional questions are prepared to delve deeper into the subject. The focus group was done in Dutch, because the research is done in the Netherlands and therefore the teachers teach in Dutch, to create a comfortable situation for discussion. The information obtained during the focus group and the questions are translated for this report. The translated sentences were translated back using automatic translators to check if the meaning was not changed during translation.

1. How is technology presently used in the classroom?
    - Which technologies are especially useful for teaching chemistry?
    - Do you see any differences in how students use these technologies?
2. Are there major differences between students in how they learn chemistry?
    - What do students prefer to use to study chemistry?
3. What are the current problems students face in writing chemistry formulas by hand?
    - Are there specific formulas that are a difficulty for students?
4. How do you grade the handwritten chemical formulas now?
5. How do you give feedback to students?
    - Do you see a difference in how students respond to feedback?
    - Which method of feedback do you estimate to be most effective?
6. How do you see the implementation of an app in the lessons?
    - In which part of the education do you see the app work?

- Should there be a differentiation for students with different learning styles and capabilities?

7. Which criteria do you think are most important when evaluating the effect of technology on the learning process?
8. Do you have any ideas or suggestions that could be a valuable aspect for the app?
9. Do you foresee any worries that students and parents have about implementing technology in classes?
10. Do you have any further questions or comments?

3.2 Focus Group Evaluation

The focus group was held with ten participants from different high schools in the east of the Netherlands. The teachers all have at least seven years of experience and teach at levels ranging from HAVO third class to VWO sixth class. Senior general secondary education (HAVO) prepares pupils for higher professional education (HBO) and takes five years to complete. Pre-university education (VWO) prepares pupils for university studies and takes six years to complete. The focus group had a time limit of thirty minutes due to conflicting schedules. Because of the time limit, questions 7-9 were not discussed, as they were deemed least important for the current project. For the evaluation, the focus group was recorded and transcribed. Since it was a focus group, not all participants answered every question, and some participants simply agreed with the statements made by other participants during certain parts of the questions and discussion.

The first question was focused on technology in the classroom. The used technologies included utility programs such as powerpoint, teams and the school system where students can see their homework, upload homework and see their grades. Another group of programs was focused on testing the knowledge of the student, such as kahoot, exit ticket and a program linked to the teaching books that they use.And lastly, the technologies used include computers, tablets, drawing tablets, and phones for utilizing the programs mentioned before. After this we talked about phones during classes, since the rules have changed recently to forbid students having a phone in the classroom and the schools handle this situation differently, from allowing the phone when the teacher knows beforehand that it will be used in the class to not allowing it at all.

From the used technologies the interview focused on how students use technologies and the differences between them.One teacher mentioned that there is a significant range in students' familiarity with the technology, from having no knowledge at all to knowing everything, and all levels in between. The other teachers were divided about this, about halve agreed and the other halve disagreed. They also observed that students can do the tasks easy on the phone, but when it comes to doing tasks on the computer, they have a lot of difficulties.

The next topic was about major differences between students in how they learn chemistry. The students need to learn a lot of formulas and be able to write them down directly. However, most of the learning takes place outside of class, and teachers are generally unaware of how students study at home. They did provide a few examples, such as using flashcards and programs like WRTS. WRTS is a program where students can input relationships similar to flashcards, and it then quizzes them in random order until all answers are correct. One response was that the final national exam in high school has to be written by hand in the Netherlands, so writing is a big part of the education during class.

Following from the previous answer that students have to write by hand was the topic of problems that student have with writing chemical formulas. The main problem teachers observe is the confusion between capital and lowercase letters, as well as the use of subscript, superscript, and regular numbers. The teachers also mentioned that the problem for subscript, superscript, and normal numbers is also difficult to learn for students, since many programs do not show or even allow these types of writing. Even the news and other places where you might find chemical formulas often ignore subscript and superscript, which makes it more difficult for students to understand this distinction.A problem teachers have noticed with capital and lowercase letters is that in the official Arial 12 font, the uppercase "I" looks identical to the lowercase "l". Teachers also mentioned a clear difference between auditory and visual students with an example Ag and Hg, *Lood* and *Jood*, which sound the same in Dutch. An overview of common mistakes can be found in table 1.

The next topic is how teachers grade the chemical formulas. The teachers said direct that it was a very gray area, since it differs from time to time. In the beginning they are more lenient, while later on they become more strict, for example in the beginning a constant mistake is only one point reduction while a year later it is a mistake for every time. Subscripts and superscript are mostly graded on relative size to the letters and its location. Capitals and lowercase letters can also be graded based on their relative size to each other. Some students write a very large first letter and a smaller capital letter as the second because, according to the teachers, they can only write in either capitals or lowercase characters, not a mix of both.

Table 1: Examples of Common Mistakes in Chemistry Formulas

Mistake	**Correct Answer**
CO (carbon monoxide)	Co (Cobalt)
h2o	H_2O
NH3	NH_3
C0	CO
H2O2	H_2O_2
So4	SO_4
H3O+	H_3O^+
CH4	CH_4
NaCl2	NaCl
Ag (Silver)	Hg (Mercury)
I (Iodine)	l (lowercase L)
Lood (Lead, Dutch)	Jood (Iodine, Dutch)

Continuing with the feedback that the teachers give to the students. The teachers said that it is based on the type of question and the mistake. They use a lot of scribbles, short notes and abbreviations. This is also depended on how far the students are in education, because one teacher said that at one point she will just give the points and let the student figure out where they made the mistake. The teachers agreed that the benefits of these methods are that students ask questions about the mistakes either to the teacher or with another student. The teachers mention that this is very useful for their development. The teachers want the students to be active in thinking about their mistakes or finding help with others.

The final topic was the implementation of the app within the educational system. For this the teachers had a couple of ideas. The first idea is to give feedback in steps, so students need to answer the question and the first form of feedback for a wrong answer should be more global, while a later feedback can be in the form of what is correct or what is something they should look for. Another idea was to show the correct answers by using green color to give an indication of what is good and what they still need to look at. And lastly the teachers mentioned that it would be good to give the students an option to choose how they would like to receive feedback, given that it has effect for learning.

The final question did not lead to new insights for this project, but it answered the questions the teachers had about what is possible and what they can expect.

4 Methodology

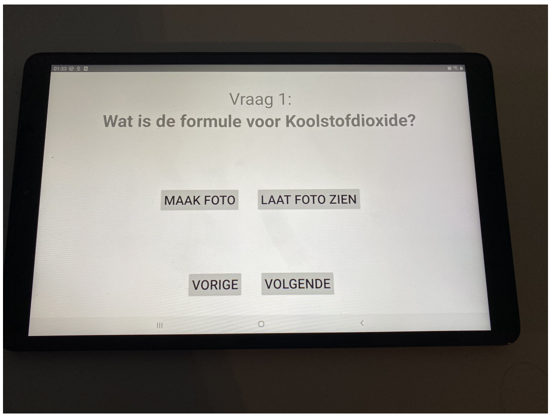
In order to understand both the input, processing and feedback methods the research is divided into three stages. The first stage focuses on testing the input, the second stage focuses on processing, while the third stage focuses on feedback. In this section a detailed explanation of the research is given.

4.1 Research Design

4.1.1 Input Stage

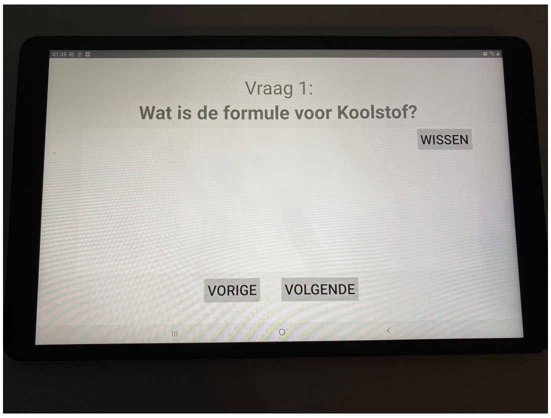
The input stage of this research focuses on evaluating three different methods for entering handwritten chemical equations into the application. The objective is to determine the most effective and user-friendly method for high school students. According to the expert focus group, these methods were selected based on their potential to enhance student engagement and accuracy in entering chemical equations. Firstly, *writing on Blank Paper*: Students write their chemical equations on plain white paper. After writing, they use the app to take a photograph of their handwritten responses. The app processes these images to recognize and analyze the handwritten chemical equations. Secondly, *writing in Grid Boxes*: In this method, students write their chemical equations within pre-drawn grid boxes on paper. The grid boxes provide structure, potentially aiding in the recognition and clarity of the handwritten text. Similar to the first method, students photograph their responses using the app for processing. Thirdly, *writing Directly on a Tablet*: Students use a stylus to write their chemical equations directly on a tablet. This method leverages the digital interface to capture the handwriting in real-time, which can then be immediately processed by the app.

The input stage was implemented using an Android application developed in Android Studio with Kotlin. The application was designed to capture images of handwritten chemical equations, irrespective of the input method. For the writing on white paper and boxed paper the question was shown with the options to take a picture and to show the picture as can be seen in 1. For the writing on the tablet method a draw field was created for the participants to write their answer as can be seen in figure 2



Vraag 1:
Wat is de formule voor Koolstofdioxide?

Figure 1: Writing on blank paper and in grid boxes: Taking pictures of written answer



Vraag 1:
Wat is de formule voor Koolstof?

Figure 2: Writing on the tablet: Canvas on the tablet

Each participant in the study went through all of the three input methods and asked to answer the same set of chemical formula questions for each method. This set consisted of two easy questions, one question about alkanes and one difficult question. The easy questions consisted of single elements or frequent formulas and difficult questions consisted of multiple elements. The questions can be seen in table 2 The app then captured their inputs for subsequent analysis. This approach was randomized to ensure that the order of methods varied among participants, minimizing any bias due to the sequence in which the methods were presented.

Table 2: Questions categorized by difficulty

Easy	**Alkanes**	**Difficult**
Water	Methane	Ethanol
Carbon dioxide	Ethane	Diphosphorus pentoxide
Cobalt	Propane	Dinitrogen tetroxide
Ammonia	Butane	Sulfur trioxide
Carbon monoxide	Pentane	Phosphorus trichloride
Chlorine	Hexane	
Oxygen	Heptane	
Carbon	Octane	
Iodine		

Following the input methods, participants were asked to complete a Technology Acceptance Model (TAM) questionnaire tailored to gather their feedback on the perceived ease of use and perceived usefulness for each input method. The TAM questionnaire was structured using the statements below, where 'method' corresponds to one of the three methods. The full questionnaire can be found in appendix A. Participants rated each statement on a 5-point Likert scale, ranging from "Strongly Disagree" to "Strongly Agree." as can be seen in figure 3.

Perceived Usefulness

1 I found it convenient to enter my answers into the app by first writing them on 'method'.

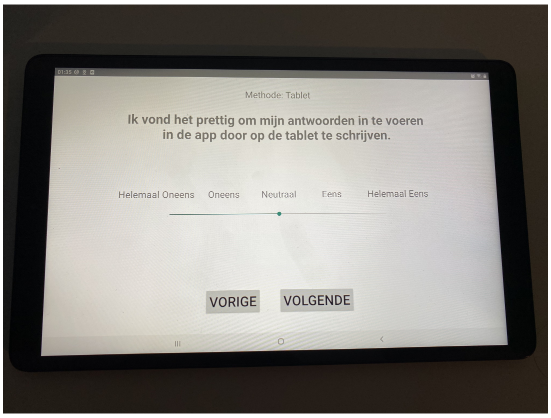
2 I would be willing to use the method of writing on 'method' again for future activities with the app.

Perceived Ease of Use

1 I found it pleasant to write my answers on 'method'.

2 It was easy to photograph my answers on 'method' using the app.

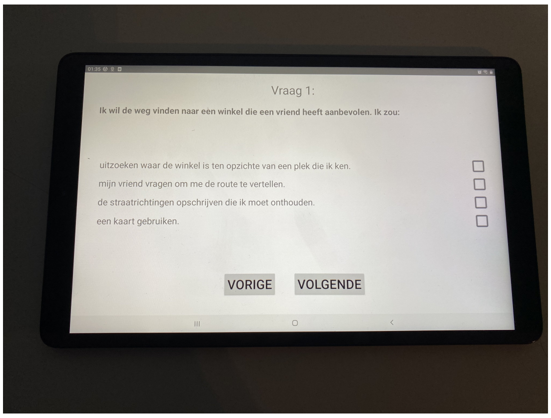
3 I enjoyed entering my answers into the app after writing them on 'method'.



Methode: Tablet

Ik vond het prettig om mijn antwoorden in te voeren in de app door op de tablet te schrijven.

Figure 3: TAM questionaire example question



Vraag 1:

Ik wil de weg vinden naar een winkel die een vriend heeft aanbevolen. Ik zou:

uitzoeken waar de winkel is ten opzichte van een plek die ik ken. □
mijn vriend vragen om me de route te vertellen. □
de straatrichtingen opschrijven die ik moet onthouden. □
een kaart gebruiken. □

Figure 4: VARK learning styles test example question

After the questionnaire, the participants completed the VARK learning styles test. This test, sourced from the official VARK website, included both the questions and an answer sheet to determine each participant's learning style. The VARK framework categorizes learning styles into four types: Visual, Aural, Read/Write, and Kinesthetic. Participants answered multiple-choice questions, selecting as many responses as they felt were true for them. An example of a question with answers can be seen in figure 4.

4.1.2 Processing Stage

In the processing stage of this research, existing datasets of images were utilized to develop a machine learning model capable of generating feedback on handwritten chemical equations. This stage involved creating several models using different datasets and preprocessing techniques for accurate recognition and feedback generation.

The initial step in the model development process involved creating a foundational model using the MNIST dataset, which contains images of handwritten digits, and the A-Z

Handwritten dataset, which includes images of handwritten letters. These datasets were loaded and combined to form a comprehensive foundation for recognizing individual characters and numbers. The MNIST dataset was preprocessed by stacking the training and test data, while the A-Z dataset involved reading and reshaping the images to 28x28 pixels.

To ensure the A-Z characters were not misclassified as digits, the labels for the A-Z dataset were offset by adding 10. Both datasets were resized to 32x32 pixels to match the input requirements of the model architecture, and the pixel values were normalized to the range [0, 1]. The combined data and labels were then split into training and testing sets for further model training.

Following the creation of the foundational model, the focus shifted to developing a model capable of recognizing handwritten chemical formulas, including common mistakes. For this, a specific dataset containing images of handwritten chemical formulas with various errors was used. The preprocessing involved resizing the images to 125x125 pixels to avoid overflow problems during training and adding a channel dimension to each image.

However, this model encountered challenges due to the limited variations in the types of mistakes. The scarcity of diverse examples for each error type hindered the model's ability to generalize and accurately identify mistakes.

To address these limitations, a third model was developed that categorized images as either correct or wrong for a selection of chemical formulas. This model required more than two labels for effective training. The preprocessing involved resizing the images to 125x125 pixels and adding an identifier for the specific question as the first row and column in the image.

The final model further refined the approach by creating a correct/wrong classification for each individual question to provide more precise feedback on specific answers. Similar preprocessing techniques were applied, including resizing the images to manageable sizes and adding a numerical identifier for the question. This model focused on providing feedback for each specific question rather than generalizing across multiple questions.

Each model underwent rigorous training and evaluation for accuracy and reliability. The models were built using a ResNet architecture, a deep residual network that helps in training very deep networks by using skip connections. The training was conducted using TensorFlow, a powerful machine learning framework. The training process included data augmentation techniques such as rotation, zoom, width and height shift, shear, and horizontal flip to enhance the robustness of the model.

The models were evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Cross-validation techniques were applied to ensure the models' performance was consistent across different subsets of the data. The models were also saved for future use.

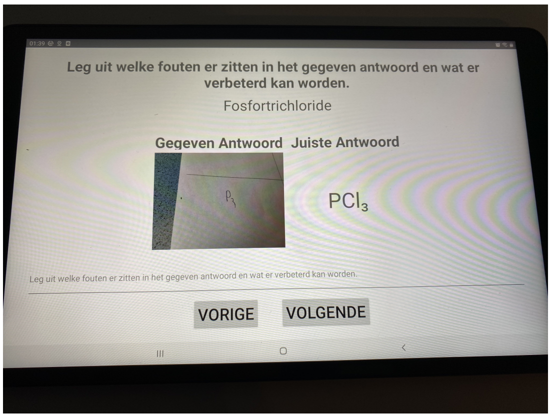
The key steps involved in the model development process included data loading and preprocessing, model training, and evaluation. Data loading and preprocessing involved combining MNIST and A-Z Handwritten datasets, resizing images to 32x32 pixels, and normalizing pixel values. For the chemical formulas, images were resized to 125x125 pixels.

Model training included developing a foundational model for character and number recognition, creating models for chemical formula recognition with a focus on identifying correct and incorrect answers, and using TensorFlow and a ResNet architecture for training. The evaluation process involved applying data augmentation techniques, evaluating models using accuracy, precision, recall, and F1-score, and using cross-validation for performance consistency.

4.1.3 Feedback Stage

The feedback stage of this research focuses on evaluating three different methods for delivering feedback to students based on their handwritten chemical equations, which will be randomly presented for each method. The objective is to determine the most effective way to provide feedback that enhances students' understanding and ability to write chemical formulas correctly.

The first method, *self-assessment with explanation*, involves presenting students with their answers alongside the correct answers. This approach prompts students to identify their mistakes and reflect on their performance, encouraging them to engage in self-assessment and develop critical thinking skills. The app then provides detailed explanations for each mistake, helping students understand the nature of their errors and how to improve. This method is shown in Figure 5. The rationale behind this method is to foster independent learning and self-correction, which can be particularly effective for students who are motivated to understand the material deeply. This method was emphasized by the expert focus group as crucial for promoting self-assessment and independent learning among students

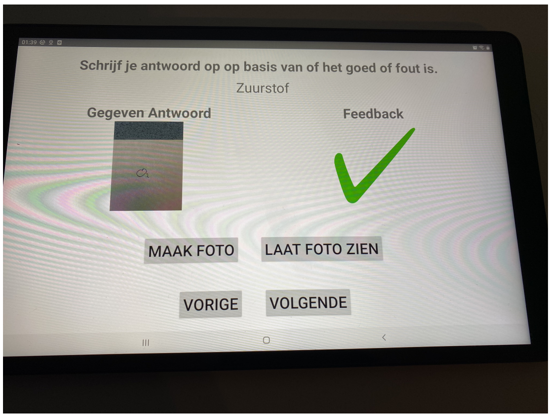


Leg uit welke fouten er zitten in het gegeven antwoord en wat er verbeterd kan worden.

Fosfortrichloride

Leg uit welke fouten er zitten in het gegeven antwoord en wat er verbeterd kan worden.

Figure 5: Showing the correct answer



Schrijf je antwoord op op basis van of het goed of fout is.

Zuurstof

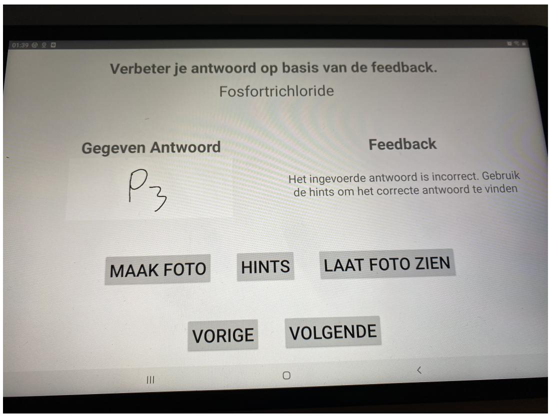
Figure 6: Showing correct or wrong

The second method, *simple marking*, displays students' answers with straightforward marks indicating correctness or incorrectness. Mistakes are highlighted with symbols such as crosses or checks. Students are then required to use these marks to correct their answers, promoting an understanding of their mistakes through minimal guidance. This method simulates the traditional feedback approach commonly used by teachers, providing a familiar context for students. The implementation of this method is depicted in Figure 6. The reasoning behind this method is to replicate the current feedback practice, allowing for a direct comparison with more advanced methods.

The third method, *detailed feedback with hints*, shows students their answers along with a comprehensive explanation of their performance. The feedback includes specific comments on what was correct and what needed improvement, along with hints to guide students in correcting their mistakes. The feedback provided through this method is tailored to the specific issues encountered by students, aiming to maximize learning by offering in-depth explanations and targeted hints. Below are the specific feedback sentences used in this method. This method aims to provide thorough explanations and guidance, as seen in Figure 7 and the hints shown in Figure 8. According to the teachers from the focus group, this is the ideal version of feedback, as it combines detailed feedback with hints, making it easier for students to understand and learn from their errors. This method is designed to maximize learning by providing in-depth explanations and targeted hints that support the learning process.

1. "It appears that the font size is incorrect. Check for uppercase and lowercase letters."
2. "The letters seem to be in the wrong place. Check for subscript, superscript, or normal position."
3. "The elements are correct, but the quantities are not yet right."

4. "The entered answer is incorrect. Use the hints to find the correct answer."
5. "It looks like you need some help with this answer. Check the available hints for information."
6. "Well done! The answer is correct, but pay attention to the font size."
7. "You have the correct answer, but pay more attention to the position for subscript and superscript."
8. "Well done! Ensure that only the answer is in the photo."
9. "You have the correct answer!"



Verbeter je antwoord op basis van de feedback.

Fosfortrichloride

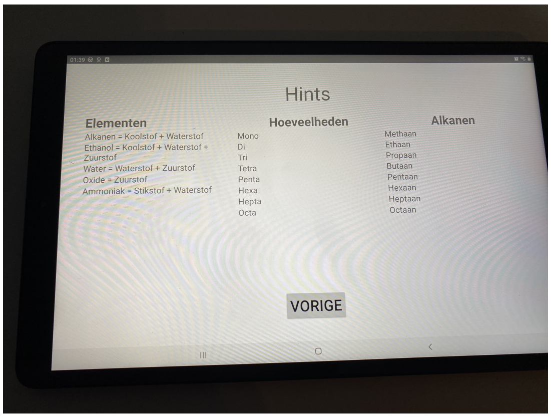
Gegeven Antwoord

P_3

Feedback

Het ingevoerde antwoord is incorrect. Gebruik de hints om het correcte antwoord te vinden

Figure 7: Showing feedback in a sentence



Hints

Elementen	Hoeveelheden	Alkanen
Alkanen = Koolstof + Waterstof	Mono	Methaan
Ethanol = Koolstof + Waterstof + Zuurstof	Di	Ethaan
Water = Waterstof + Zuurstof	Tri	Propaan
Oxide = Zuurstof	Tetra	Butaan
Ammoniak = Stikstof + Waterstof	Penta	Pentaan
	Hexa	Hexaan
	Hepta	Heptaan
	Octa	Octaan

Figure 8: Hints for the questions

The feedback stage is setup in the same way as the input stage using android studio to create an app that guides participants through the different stages. The order of methods is again randomized to minimize any bias.

Following the feedback methods, participants were asked to again complete a Technology Acceptance Model (TAM) questionnaire to gather their feedback on the ease of use and perceived usefulness for each feedback method. The TAM questionnaire was constructed using the statements seen below where 'method' is replaced with each of the three methods. The complete questionnaire can be found in appendix A. Participants rated each statement on a 5-point Likert scale, ranging from "Strongly Disagree" to "Strongly Agree." Once the participants were done they were shown all the questions from table 2 with the correct answers.

Perceived Usefulness

1. I find ’method’ useful for receiving feedback.
2. ’Method’ helps me understand the feedback better.
3. I find ’method’ useful for learning chemical formulas.
4. I find ’method’ effective for learning chemical formulas.

Perceived Ease of Use

1. ’Method’ is easy to use.
2. ’Method’ is clear.
3. I quickly mastered the tool for ’method’.
4. I find it easy to improve in chemistry using ’method’.
5. I feel comfortable when using ’method’.

4.2 Participants

The research participants were drawn from high schools in the vicinity of Enschede, and ranged in age from 13 to 16. The study focused on students with nearly one year of experience in Havo 3 or Vwo 3, which corresponds to the third year of high school in the Dutch education system. This initial pool of participants was expected to experience six possible orders of experimental setups. The target number of participants was approximately ten students per experiment, totaling around 60 students. Participants will take part in the first experiment about input methods and in the third experiment about feedback mechanisms.

4.3 Data collection

Data collection in this study involves a systematic approach to ensure that all relevant information about participant interactions with the app is captured accurately. The process begins with participants logging into the application using a unique identifier. This login screen is designed to collect demographic information such as age, gender, school, and student number, which helps in understanding the diversity of the participant pool. During each session, various types of information are collected and stored.

Firstly, participant information is recorded. This includes age, which is noted to analyze performance and preferences across different age groups, and gender, to examine any gender-based differences in interaction and feedback. The gender options follow the guidelines set in Spiel et al. (2019) to make sure that everyone feels included. The school information is collected to differentiate between different educational environments, and each participant is using their schools student number as a unique identifier to track individual performance while maintaining confidentiality.

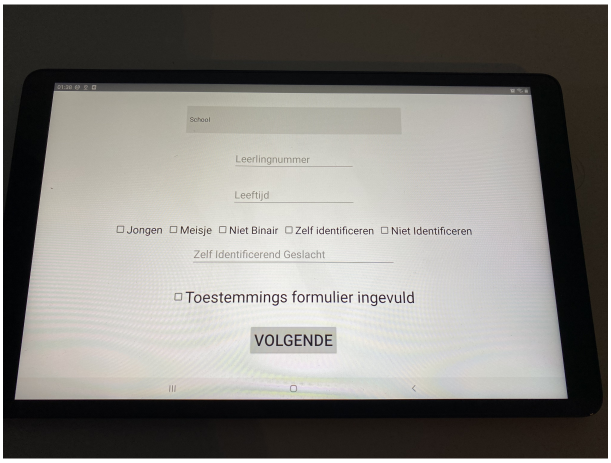
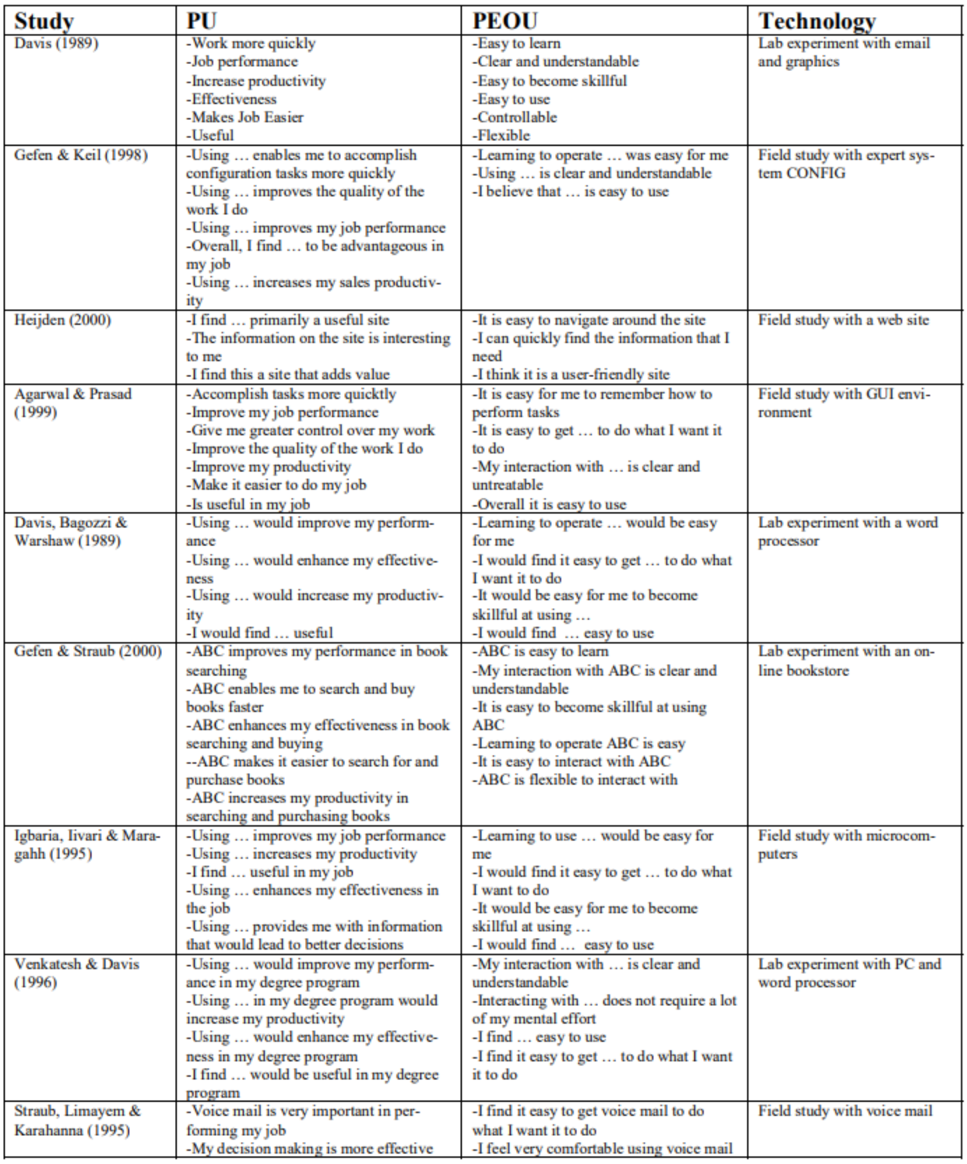

School
Leerlingnummer
Leeftijd
□ Jongen □ Meisje □ Niet Binair □ Zelf identificeren □ Niet Identificeren
Zelf Identificerend Geslacht
□ Toestemmings formulier ingevuld
VOLGENDE

Figure 9: Login screen

Interaction data is also meticulously gathered. This includes the order of methods, which details the sequence in which the different methods are presented to each participant The specific questions given to each participant, categorized by difficulty level, are also recorded. Additionally, captured images of handwritten responses when using paper-based methods are saved for OCR processing. For the tablet method, screen captures are used to safe the answers. All the answers provided by the participants are stored in text format for subsequent analysis.

Feedback data is another crucial aspect of the collection process. After the input methods session, participants complete a questionnaire based on the Technology Acceptance Model (TAM) Davis and Davis (1989). This includes ratings for ease of use, perceived usefulness, and overall satisfaction with each input method, aligning with the findings of Hamidi and Chavoshi (2018), Masrek and Samadi (2017), and Sarrab et al. (2016), who highlighted the importance of these factors in an educational context. Additionally, Ma and Liu (2006) illustrates how TAM is employed in similar research, as shown in 10. The TAM questions used for the input stage and for the feedback stage can be found in appendix A.



Study	PU	PEOU	Technology
Davis (1989)	-Work more quickly -Job performance -Increase productivity -Effectiveness -Makes Job Easier -Useful	-Easy to learn -Clear and understandable -Easy to become skillful -Easy to use -Controllable -Flexible	Lab experiment with email and graphics
Gefen & Keil (1998)	-Using … enables me to accomplish configuration tasks more quickly -Using … improves the quality of the work I do -Using … improves my job performance -Overall, I find … to be advantageous in my job -Using … increases my sales productivity	-Learning to operate … was easy for me -Using … is clear and understandable -I believe that … is easy to use	Field study with expert system CONFIG
Heijden (2000)	-I find … primarily a useful site -The information on the site is interesting to me -I find this a site that adds value	-It is easy to navigate around the site -I can quickly find the information that I need -I think it is a user-friendly site	Field study with a web site
Agarwal & Prasad (1999)	-Accomplish tasks more quicktly -Improve my job performance -Give me greater control over my work -Improve the quality of the work I do -Improve my productivity -Make it easier to do my job -Is useful in my job	-It is easy for me to remember how to perform tasks -It is easy to get … to do what I want it to do -My interaction with … is clear and untreatable -Overall it is easy to use	Field study with GUI environment
Davis, Bagozzi & Warshaw (1989)	-Using … would improve my performance -Using … would enhance my effectiveness -Using … would increase my productivity -I would find … useful	-Learning to operate … would be easy for me -I would find it easy to get … to do what I want it to do -It would be easy for me to become skillful at using … -I would find … easy to use	Lab experiment with a word processor
Gefen & Straub (2000)	-ABC improves my performance in book searching -ABC enables me to search and buy books faster -ABC enhances my effectiveness in book searching and buying --ABC makes it easier to search for and purchase books -ABC increases my productivity in searching and purchasing books	-ABC is easy to learn -My interaction with ABC is clear and understandable -It is easy to become skillful at using ABC -Learning to operate ABC is easy -It is easy to interact with ABC -ABC is flexible to interact with	Lab experiment with an online bookstore
Igbaria, Iivari & Maragahh (1995)	-Using … improves my job performance -Using … increases my productivity -I find … useful in my job -Using … enhances my effectiveness in the job -Using … provides me with information that would lead to better decisions	-Learning to use … would be easy for me -I would find it easy to get … to do what I want to do -It would be easy for me to become skillful at using … -I would find … easy to use	Field study with microcomputers
Venkatesh & Davis (1996)	-Using … would improve my performance in my degree program -Using … in my degree program would increase my productivity -Using … would enhance my effectiveness in my degree program -I find … would be useful in my degree program	-My interaction with … is clear and understandable -Interacting with … does not require a lot of my mental effort -I find … easy to use -I find it easy to get … to do what I want it to do	Lab experiment with PC and word processor
Straub, Limayem & Karahanna (1995)	-Voice mail is very important in performing my job -My decision making is more effective	-I find it easy to get voice mail to do what I want it to do -I feel very comfortable using voice mail	Field study with voice mail

Figure 10: Example of TAM questions from Ma and Liu (2006)

Furthermore, participants complete the VARK learning styles test Fleming (1995), which helps in categorizing their preferred learning methods (Visual, Aural, Read/Write, Kinesthetic). This approach not only provides insights into the effectiveness of the input methods but also aligns with the related works' emphasis on ease of use and perceived usefulness as critical factors for successful technology integration in education.

The collected data is stored in CSV files for ease of processing and analysis. Each CSV file includes columns for participant ID, demographic information, the order of input methods, specific questions, captured images, screen captures, answers provided, and questionnaire responses. This structured format allows for efficient data handling and subsequent statistical analysis.

4.4 Data Analysis

The data analysis for this study was designed to evaluate the effectiveness of different input and feedback methods for learning chemical formulas, as well as to assess students' preferences and learning styles. The analysis was conducted using both quantitative and qualitative methods to provide a comprehensive understanding of the collected data.

Demographic data, including age, gender, school, and student number, were first analyzed to provide a baseline understanding of the participant population. Descriptive statistics were used to summarize the demographic characteristics of the participants.

The accuracy of students' answers to chemical formula questions was assessed across the different input methods. For each participant, the correctness of their responses was recorded and compared across the three input methods: writing on plain paper, writing in grid boxes, and writing directly on the tablet. This analysis helped identify which method yielded the highest accuracy in student responses.

To determine students' perceptions of ease of use and perceived usefulness for the input and feedback methods, responses to the Technology Acceptance Model (TAM) questionnaire were analyzed. The TAM questionnaire consisted of Likert scale statements, rated by participants on a scale from 1 (strongly disagree) to 5 (strongly agree). These statements captured various aspects of their experiences with the methods.

The analysis process began with a factor analysis of the TAM questions. This step was crucial for understanding the underlying structure of the questionnaire and identifying the main factors contributing to perceived ease of use and perceived usefulness. By performing the factor analysis, we aimed to determine which questions were most significant in measuring these constructs. Following the factor analysis, Cronbach's alpha was calculated for each combination of methods with perceived ease of use and perceived usefulness. Cronbach's alpha is a measure of reliability that assesses the degree to which a set of items measures a single unidimensional latent construct. A higher alpha value indicates greater internal consistency among the items. In this study, Cronbach's alpha was used to verify the reliability of the TAM questionnaire responses.

Next, the average TAM scores were calculated for each method. This involved computing the mean responses for perceived ease of use and perceived usefulness separately for each method. By comparing these averages, we could identify trends and preferences among the participants. To ensure the statistical significance of these differences, an ANOVA test was conducted for each method's TAM scores. This test helped determine whether the observed differences in TAM scores among the input methods were statistically significant.

Additionally, the influence of the order in which the methods were presented was analyzed. This involved examining whether the sequence in which participants experienced the methods affected their TAM scores. Specifically, we looked at how starting with different methods influenced the subsequent ratings of perceived usefulness and ease of use for the other methods.

To further refine our analysis, we examined the impact of participant demographics, including age, gender, and the school attended, on the TAM scores. By analyzing these factors, we aimed to understand whether different age groups, genders, or schools influenced participants' perceptions of perceived usefulness and ease of use for the input and feedback methods. This comprehensive approach allowed us to identify any notable variations in experiences and provided a deeper understanding of the strengths and potential areas for improvement in each method.

The VARK questionnaire responses were analyzed to categorize students into four distinct learning styles: Visual, Aural, Read/Write, and Kinesthetic. Each participant's responses to the VARK questionnaire were scored according to the guidelines provided by the VARK framework. These scores were then used to classify each student into one of the four learning styles based on their highest score. Once the learning styles were determined, the distribution of these styles among the participants was examined to understand the overall makeup of the study group. This step was crucial in ensuring that all learning styles were adequately represented in the analysis.

The next phase involved exploring the correlations between learning styles and preferences for the different input and feedback methods. This was achieved by comparing the TAM scores for perceived ease of use and perceived usefulness across the different learning styles. For each learning style category, the average TAM scores were calculated and analyzed to identify any significant trends or preferences. By integrating the VARK learning styles with the TAM model, we aimed to uncover whether certain input and feedback methods were more effective or preferred by students with specific learning styles. This analysis provided insights into how different types of learners interact with various educational tools and helped identify which methods might require adjustments to better cater to the diverse needs of students. The statistical significance of these relationships was determined using ANOVA tests.

Observational data collected during the study sessions were analyzed to gain insights into student interactions, emotions, and engagement with the different input and feedback methods.

The results from the quantitative and qualitative analyses were integrated to provide a comprehensive understanding of the study findings. This integration helped triangulate the data, ensuring that the conclusions drawn were robust and supported by multiple sources of evidence.

Results were visualized using charts, graphs, and tables to facilitate interpretation and presentation. Bar charts were used to illustrate demographic distributions, while bar graphs and line charts were employed to compare performance and preferences across different methods. Visualizations helped in clearly communicating the findings and making data-driven conclusions. The statistical results from the ANOVA tests were also presented to provide a robust understanding of the significance of the observed differences and trends.

4.5 Ethical considerations

Ensuring ethical conduct throughout the research process is of great importance. This study has been reviewed and approved by the Ethics Committee Information and Computer Science at the University of Twente, ensuring that all procedures meet ethical standards and protect the rights and well-being of participants.

Informed Consent: Prior to participation, informed consent was obtained from all students involved in the study. This included providing detailed information about the purpose of the research, the procedures involved, the potential benefits and risks of participation, and the voluntary nature of their involvement. Additionally, parental consent was obtained for participants under the age of 16, ensuring that all guardians were fully informed and in agreement with their child's participation in the study.

Confidentiality and Privacy: Participants' personal information, such as age, gender, school, and student number, was collected solely for the purpose of analyzing demographic data and ensuring the integrity of the research. This information is stored securely and is only accessible to the research team. To protect privacy, data was anonymized by assigning unique identifiers to each participant, ensuring that individual responses could not be traced back to specific participants. No personal identifiers were used in the analysis or presentation of the research findings.

Data Security: All data collected during the study, including demographic information, interaction data, and questionnaire responses, is securely stored. Physical copies of handwritten answers and any other sensitive information are kept securely. Digital data is stored on password-protected computers and secure servers. Data will be retained for a period of five years following the completion of the study, after which it will be securely destroyed.

Voluntary Participation and Withdrawal: Participation in the study was entirely voluntary. Participants were informed that they could withdraw from the study at any time without any consequences. If a participant chose to withdraw, their data would be excluded from the study and securely destroyed. This ensured that participants felt comfortable and had full control over their involvement in the research.

Minimizing Risks and Discomfort: While the study posed minimal risk to participants, measures were taken to minimize any potential discomfort. Participants engaged in familiar educational tasks, albeit through an app interface, and were provided with clear instructions and support throughout the process. Any potential discomfort from engaging in unfamiliar tasks was mitigated by the presence of researchers to assist and answer questions as needed.

Transparency and Communication: Throughout the study, clear communication was maintained with participants and their guardians. Any questions or concerns were addressed promptly, ensuring that participants felt informed and supported. Contact information for the researcher and the Ethics Committee was provided to all participants, allowing them to seek further information or report any issues related to the study.

By adhering to these ethical considerations, the research aims to uphold the highest standards of integrity and respect for all participants, ensuring that their rights and well-being are prioritized throughout the study.

5 Results

5.1 Participants

In the end the study included about 160 students from seven different classes. However, only 59 of these students submitted the required ethics consent form, which allowed their data to be included in the final analysis.

To provide a clearer demographic breakdown, the following charts 11, 12 and 13 illustrate the age, learning styles and gender distribution of the participating students. For the learning styles only the main learning style is shown. This demographic information helps in understanding the context of the participants and in assessing the generalizability of the study's findings.

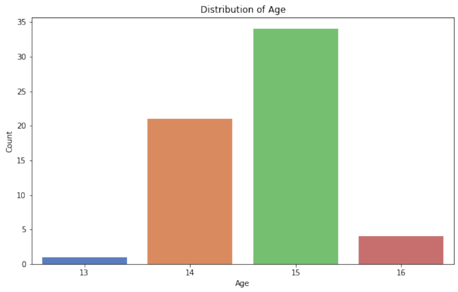


Figure 11: Age spread of participants

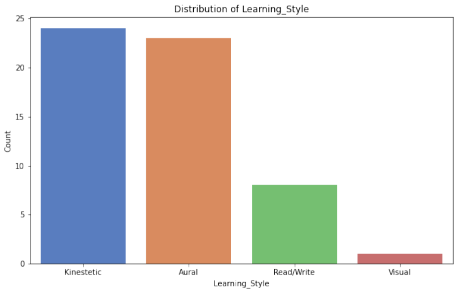


Figure 12: Learning style spread of participants

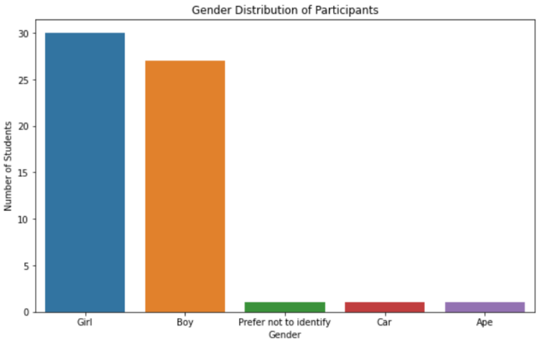


Figure 13: Gender spread of participants

The age spread reveals that the average age of participants is either 14 or 15, with a few students being 13 or 16. This indicates that the majority of students are in the mid-teen age range, which is typical for high school students.

When examining the gender distribution, it can be seen that most participants identify themselves as either a boy or a girl. However, there are two outliers who identify as a car or an ape, and one participant who prefers not to identify. These responses highlight questionable answers to our attempt of being inclusive and properly following non-binary options in demographic surveys for high school students.

The learning style distribution shows that most participants fall into the Kinesthetic and Aural categories. There is only one student identified as Visual, and approximately one-sixth of the participants are classified as Read/Write learners.

In summary, the majority of participants fit within common demographic categories. However, the low numbers in certain groups (ages 13 and 16, Visual learning style, and non-binary gender identifications such as car and ape and prefering not to identify) mean that the results related to these groups should be interpreted with caution regarding their generalizability.

5.2 Input Stage

Due to an error in the app, the written answers for writing directly on the tablet were not saved correctly for two-thirds of the participants. Despite this, the analysis of the remaining data indicates that the method used did not significantly impact the number of correct answers provided by the participants, nor did it influence the type of feedback related to their answers.

The feedback for each method is categorized based on the type of questions. "Easy" questions include simple and familiar compounds, "Alkanes" questions involve hydrocarbons from methane to octane, and "Difficult" questions are those that are less common. This categorization helps in understanding the type of feedback provided and its distribution, as shown in the figures 14, 15 and 16.

Analyzing the feedback distribution reveals that for "Easy" questions, there was a high frequency of correct answers and fewer instances of incorrect quantities, size, or placement. For "Alkanes" questions, the feedback predominantly indicated incorrect answers, reflecting the students' lack of familiarity with these compounds. For "Difficult" questions, the feedback was varied, showing instances of incorrect quantities, placement, size, and some "no idea" responses, highlighting the complexity and unfamiliarity of these questions to the students.

Analyzing the responses from the Technology Acceptance Model (TAM) questionnaire provided insightful findings regarding the factor structure and internal consistency of the items related to different methods of writing.

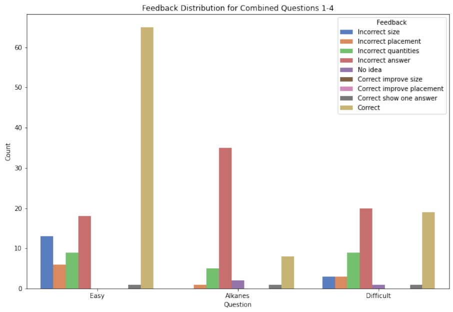


Figure 14: Feedback distribution for writing on blank paper

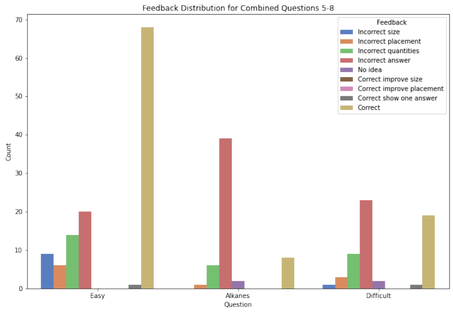


Figure 15: Feedback distribution for writing in grid boxes

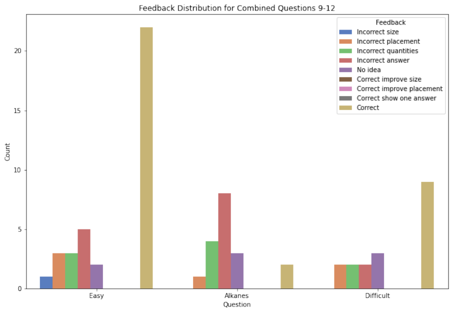


Figure 16: Feedback distribution for writing on a tablet

Factor analysis was conducted to examine the underlying structure of the TAM questions for each writing method: blank paper, grid boxes, and tablet writing. For the blank paper method, the factor loadings indicated that questions 3 and 4 had higher loadings on Factor 1, suggesting these items were more aligned with perceived usefulness. In contrast, questions 2 and 5 showed moderate loadings on both factors, reflecting a mix of perceived ease of use and usefulness, while question 1 loaded very low on factor 1 and moderately on factor 2. The grid boxes method revealed that questions 8 and 9 had higher loadings on Factor 1, indicating their strong association with perceived usefulness, while questions 7 and 10 loaded moderately on both factors and question 6 loading very low on factor 1 and moderate of factor 2. For tablet writing, questions 13 and 14 showed significant loadings on Factor 1, emphasizing their relevance to perceived usefulness, while questions 11 and 12 indicated a mix of ease of use and usefulness. See table3 for the full results.

To ensure the reliability of the TAM questionnaire, Cronbach's alpha was calculated for each combination of methods with perceived ease of use and perceived usefulness. For the blank paper method, the Cronbach's alpha values were 0.46 for ease of use and 0.65 for usefulness, demonstrating low internal consistency for ease of use and moderate internal consistency for usefulness. For the grid boxes method, the Cronbach's alpha values were

Table 3: Factor Loadings for Different Writing Methods

TAM Question	Factor 1	Factor 2
Blank Paper		
Q1	0.056	0.32
Q2	0.30	0.45
Q3	0.70	0.49
Q4	0.98	0.17
Q5	0.40	0.40
Grid Boxes		
Q6	0.084	0.35
Q7	0.17	0.45
Q8	0.68	0.55
Q9	0.99	0.16
Q10	0.54	0.42
Tablet Writing		
Q11	0.22	0.97
Q12	0.52	0.14
Q13	0.84	0.10
Q14	0.80	0.17
Q15	0.76	0.24

0.39 for ease of use and 0.74 for usefulness, indicating low internal consistency for ease of use and high internal consistency for usefulness. For the tablet writing method, the Cronbach's alpha values were 0.61 for ease of use and 0.80 for usefulness, meaning a moderate internal consistency for ease of use and a high internal consistency for usefullness. See table 4 for the full results.

Table 4: Cronbach's Alpha for Different Writing Methods

Writing Method	Ease of Use (Alpha)	Usefulness (Alpha)
Blank Paper	0.46	0.65
Grid Boxes	0.39	0.74
Tablet Writing	0.61	0.80

The TAM questionnaire results necessitate careful interpretation, as most Cronbach's alpha scores are below 0.7. This indicates potential issues with the internal consistency of the questionnaire. Despite this, the factor loadings do provide some insights into students' perceptions of ease of use and usefulness across different writing methods, highlighting the importance of specific questions in capturing these perceptions and offering valuable information for further analysis and potential educational practice improvements.

Analyzing the responses from the TAM questionnaire reveals several key insights. The method of writing (blank paper, grid boxes, or tablet writing) did significantly impact the overall TAM scores. This was supported by ANOVA results, which showed significant differences for ease of use (F(2, 56) = 34.780, p = 1.964e-13) and usefulness (F(2, 56) = 17.798, p = 9.278e-08). Further analysis using Tukey's HSD test provided a detailed comparison between the methods. For ease of use, the comparison revealed that tablet writing was significantly preferred over blank paper (mean difference = 1.05, $p < 0.001$) and grid boxes (mean difference = 0.90, $p < 0.001$), whereas there was no significant difference between blank paper and grid boxes (mean difference = 0.15, p = 0.50). For usefulness, tablet writing was also significantly preferred over blank paper (mean difference = 0.93, $p < 0.001$) and grid boxes (mean difference = 0.80, $p < 0.001$), with no significant difference between blank paper and grid boxes (mean difference = 0.14, p = 0.70). These results indicate that participants had a strong preference for tablet writing over both blank paper and grid boxes based on ease of use and perceived usefulness.

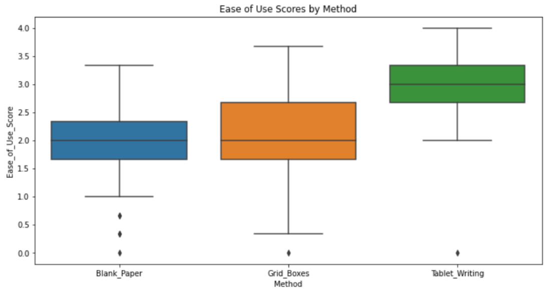


Figure 17: Ease of Use Scores by Method

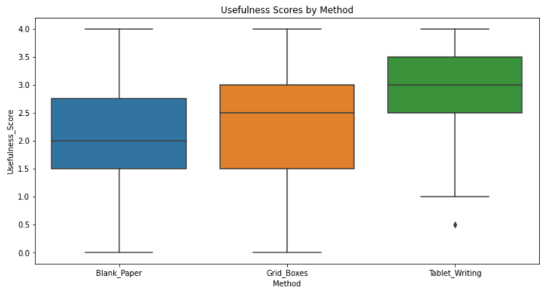


Figure 18: Usefulness Scores by Method

Table 5: Tukey's HSD Results for Writing Methods

Metric	Group 1	Group 2	Mean Diff.	p-adj	Significant
Ease of Use	Blank Paper	Grid Boxes	0.1525	0.5029	No
Ease of Use	Blank Paper	Tablet Writing	1.0508	0.0000	Yes
Ease of Use	Grid Boxes	Tablet Writing	0.8983	0.0000	Yes
Usefulness	Blank Paper	Grid Boxes	0.1356	0.7016	No
Usefulness	Blank Paper	Tablet Writing	0.9322	0.0000	Yes
Usefulness	Grid Boxes	Tablet Writing	0.7966	0.0000	Yes

Table 6: ANOVA Results for Ease of Use and Usefulness by Demographic Factors

Factor	Ease of Use	Usefulness
Gender	F(1, 57) = 0.20, p = 0.66	F = 0.55, p = 0.46
School	F(2, 56) = 1.31, p = 0.28	F = 12.78, p = 2.69e-05
Age	F(3, 55) = 0.78, p = 0.46	F = 0.76, p = 0.47
Learning style	F(3, 55) = 0.78, p = 0.46	F = 0.76, p = 0.47
Experiment order	F(5, 53) = 1.26, p = 0.29	F = 2.59, p = 0.036

The impact of demographic factors on TAM scores was also examined. This analysis can be found in Table 6. Age had no significant impact on TAM scores for either Ease of Use (F(3, 55) = 0.78, p = 0.46) or Usefulness (F(3, 55) = 0.76, p = 0.47). Gender also showed no significant impact on TAM scores for either Ease of Use (F(1, 57) = 0.20, p = 0.66) or Usefulness (F(1, 57) = 0.55, p = 0.46). However, the school attended by participants did have a significant impact on TAM scores for Usefulness (F(2, 56) = 12.78, p < 0.001), while no significant impact was found for Ease of Use (F(2, 56) = 1.31, p = 0.28). Further analysis using Tukey's HSD test for school revealed that the significant difference found in the ANOVA test for Usefulness was specifically between School A and School B (mean difference = -1.20, p < 0.001) and between School A and School C (mean difference = -0.95, p = 0.001). There was no significant difference between School B and School C (mean difference = 0.25, p = 0.63), as seen in Table 5. This suggests that factors such as teaching methods, school resources, or the overall learning environment in School A might differ significantly, influencing students' perceptions of usefulness compared to the other schools.

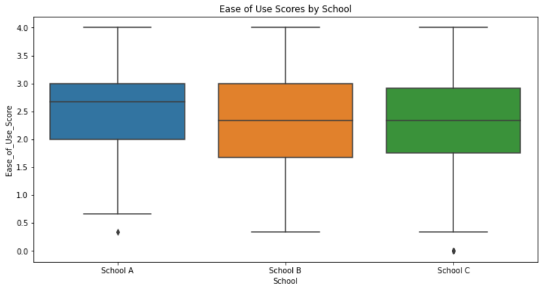


Figure 19: Ease of Use Scores by School

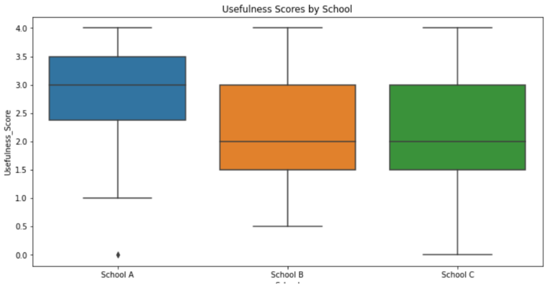


Figure 20: Usefulness Scores by School

Factor	Group 1	Group 2	Mean Diff.	p-adj	Significant
Ease of Use	School A	School B	0.13	0.6043	No
Ease of Use	School A	School C	0.09	0.7873	No
Ease of Use	School B	School C	-0.04	0.9601	No
Usefulness	School A	School B	0.15	0.4329	No
Usefulness	School A	School C	0.23	0.1645	No
Usefulness	School B	School C	0.07	0.8545	No

Table 7: Tukey HSD Results for Significant Factors (School)

The order in which the methods were presented to participants did not significantly affect ease of use scores (F(5, 53) = 1.26, p = 0.29) or usefulness scores (F(5, 53) = 2.59, p = 0.036). Learning styles did not show a significant impact on ease of use scores (F(3, 55) = 0.78, p = 0.46) or usefulness scores (F(3, 55) = 0.76, p = 0.47).

In summary, the input method did not significantly impact the correctness of students' answers, but the direct tablet input method was preferred due to its perceived ease of use and engaging nature. The analysis highlights the importance of both familiarity with the content and the appeal of modern technology in enhancing the learning experience. Additionally, the school attended by participants had a significant impact on TAM scores, suggesting that factors such as teaching methods, school resources, or the overall learning environment in School A might differ significantly, influencing students' perceptions of usefulness compared to the other schools.

5.3 Processing stage

The processing stage of the results section reveals various performance levels across different models. The model created using the MNIST and A-Z dataset achieved an impressive accuracy of 97% as can be seen in table 8 and a full analysis of each character can be found in appendix B.1. However, it is important to note that this model is not directly applicable to chemical formulas. To make it functional for chemical formulas, a system needs to be developed that can identify individual letters and numbers, incorporating some information about their size and position.

Table 8: MNIST and A-Z trained model results

Class	Precision	Recall	F1-score	Support
Accuracy	0.97	0.97	0.97	88491
Macro avg	0.96	0.96	0.96	88491
Weighted avg	0.97	0.97	0.97	88491

The second model, which used labels for all possible answers, scored an accuracy of only 0.0% with only a few correct assumptions as can be seen in table 9. This low accuracy can be attributed to the need for a large number of labels to accurately identify all possible mistakes. The dataset used did not contain enough data to adequately represent this wide range of possibilities, leading to poor performance.

Table 9: Mistakes science formulas per mistake

Class	Precision	Recall	F1-Score	Support
Accuracy	0.00	0.00	0.00	1968
Macro avg	0.00	0.02	0.00	1968
Weighted avg	0.00	0.00	0.00	1968

The third model, designed to determine whether answers were correct or wrong, achieved an accuracy of 38%. This model showed proficiency in identifying common or easy questions but struggled with more difficult questions. Its moderate success suggests that while it can handle straightforward tasks, it requires further refinement to tackle more complex problems.

Table 10: Mistakes science formulas per page

Class	**Precision**	**Recall**	**F1-score**	**Support**
Accuracy	0.38 (2013)			
Macro avg	0.26	0.20	0.18	2013
Weighted avg	0.40	0.38	0.34	2013

The final model, which evaluated the correctness of all questions individually, recorded an accuracy of 15% which can be seen in table 11. This lower accuracy could be due to the need for specific information to determine if a question is correct or wrong. Unlike the previous model, which could identify correct answers by their similarity and found it challenging to pinpoint mistakes, this model required precise identification, resulting in less reliable performance.

Table 11: Mistakes science formulas per question

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Accuracy			0.15	2013
Macro Avg	0.07	0.10	0.07	2013
Weighted Avg	0.27	0.15	0.17	2013

Overall, these results highlight the varying degrees of success in creating an OCR system for chemical formulas, emphasizing the need for more data and refined techniques to improve accuracy, especially for more complex or varied questions.

5.4 Feedback stage

There were some issues during the experiment where the app crashed and participants were shown a given answer that was correct, but the feedback indicated that it was wrong. This discrepancy could be caused by a mismatch between the feedback, question, and image or a misunderstanding of the given answer.

For the first method, where participants had to identify what they had done wrong, the responses fell into several distinct categories. Some examples for each category can be seen in table 12. Many participants struggled with correctly positioning subscripts and superscripts. Common issues included incorrect positioning of numbers, such as placing them above instead of below the symbols, and missing subscripts altogether. Another frequent issue was the incorrect use of uppercase and lowercase letters in chemical symbols. Participants often made mistakes by either incorrectly capitalizing the letters or mixing uppercase and lowercase incorrectly. For instance, they would write an uppercase 'C' when it should have been lowercase, or vice versa.

Chemical formula errors were also prevalent. These errors ranged from minor mistakes in the chemical compositions to completely incorrect formulas. Participants often had the wrong quantities of elements in their formulas or entirely incorrect chemical compositions. This indicates a misunderstanding of basic chemical naming conventions and the principles of balancing chemical equations. Additionally, many participants acknowledged they did not know the correct answer or provided no answer at all. This lack of response highlighted significant gaps in their knowledge. Some admitted they didn't know the answers, while others simply left the answer blank, indicating areas where further instruction is needed. Finally, there were miscellaneous comments that provided additional context or personal reflections. These included general remarks about needing to learn more and personal reflections on the task or their mistakes.

Table 12: Example Comments from the Feedback Stage

Category	**Comments**
Subscript and Superscript Problems	• "O is too large" • "2 is above" • "There should be a 2 behind Cl"
Capitalization Errors	• "The C must be uppercase and the o must be lowercase" • "The O must be lowercase"
Chemical Formula Errors	• "Not enough carbon, too much hydrogen and oxygen" • "It should be CH4 not NH4"
No Answer or Incorrect Answer	• "I didn't know the answer" • "No answer"
Miscellaneous Comments	• "The answer is correct" • "I need to learn more" • "Beautiful pen"

For the Correct Answer method, the factor loadings revealed several interesting patterns. Questions 1 and 2 had notably high loadings on Factor 1 (0.90 and 0.91, respectively), indicating these items strongly relate to perceived usefulness. Similarly, questions 3 through 5 also showed substantial loadings on Factor 1, suggesting they align with perceived usefulness. In contrast, questions 7 through 9 exhibited higher loadings on Factor 2 (0.82, 0.69, and 0.66, respectively), indicating these questions capture elements of perceived ease of use more effectively. And finally question 6 has an almost equal loading for both factors.

The Correct/Wrong Marks method demonstrated a clear distinction between the two factors. Questions 10 through 13 had higher loadings on Factor 2, particularly question 12 with a loading of 0.86, emphasizing their strong association with perceived usefulness. Conversely, questions 14 through 18 showed significant loadings on Factor 1, with questions 14 and 16 having high loadings of 0.88 and 0.86, respectively. This highlights their relevance to perceived ease of use.

For Detailed Feedback, the factor analysis results were equally revealing. Questions 19 and 20 showed substantial loadings on Factor 2, particularly question 20 with a loading of 0.88, underscoring their alignment with perceived usefulness. Questions 22 and 25 through 27 had high loadings on factor 1, with question 26 standing out with a high loading of 0.96, indicating a strong association with perceived usefulness, while questions 21, 23 and 24 had equal loadings on both factors, reflecting a balanced capture of ease of use and usefulness. See Table 13 for the full results.

To ensure the reliability of the TAM questionnaire, Cronbach's alpha was calculated for each combination of methods with perceived ease of use and perceived usefulness. For the Correct Answer method, the Cronbach's alpha values were 0.95 for ease of use and 0.91 for usefulness, demonstrating high internal consistency. For the Correct/Wrong Marks method, the Cronbach's alpha values were 0.92 for ease of use and 0.95 for usefulness, indicating excellent internal consistency. For the Detailed Feedback method, the Cronbach's alpha values were 0.90 for ease of use and 0.91 for usefulness, both reflecting good internal consistency. See Table 14 for the full results.

The TAM questionnaire has proven to be a reliable tool for assessing students' perceptions of ease of use and usefulness across different feedback methods, as evidenced by the factor loadings and Cronbach's alpha values. These results highlight the importance of specific questions in capturing these perceptions, providing valuable insights for further analysis and educational practice improvements.

The ANOVA results for the feedback methods revealed no significant differences in perceived ease of use or usefulness among the three methods: Correct Answer, Correct/Wrong Marks, and Detailed Feedback. For ease of use, the results were F(2, 56) = 0.64, p = 0.53, indicating that students found all methods comparably easy to use. Similarly, for perceived usefulness, the results were F(2, 56) = 0.09, p = 0.91, suggesting that none of the feedback methods stood out as significantly more useful. These findings imply that students perceive all three feedback methods to be similarly effective, with no method being distinctly easier to use or more useful than the others.

Table 13: Factor Loadings for Different Feedback Methods

TAM Question	Factor 1	Factor 2
Correct Answer		
Q1	0.90	0.27
Q2	0.91	0.31
Q3	0.76	0.44
Q4	0.73	0.56
Q5	0.76	0.48
Q6	0.57	0.53
Q7	0.20	0.82
Q8	0.56	0.69
Q9	0.55	0.66
Correct/Wrong Marks		
Q10	0.37	0.62
Q11	0.57	0.73
Q12	0.31	0.86
Q13	0.51	0.75
Q14	0.88	0.40
Q15	0.80	0.47
Q16	0.86	0.34
Q17	0.70	0.49
Q18	0.69	0.47
Detailed Feedback		
Q19	0.25	0.68
Q20	0.45	0.88
Q21	0.64	0.66
Q22	0.66	0.47
Q23	0.68	0.63
Q24	0.50	0.50
Q25	0.74	0.40
Q26	0.96	0.30
Q27	0.57	0.38

Table 14: Cronbach's Alpha for Different Feedback Methods

Feedback Method	Ease of Use (Alpha)	Usefulness (Alpha)
Correct Answer	0.95	0.91
Correct/Wrong Marks	0.92	0.95
Detailed Feedback	0.90	0.91

Here is the updated table with the Total TAM Score removed and the correct degrees of freedom accounted for:

The order in which the feedback methods were presented had a significant influence on perceived ease of use, F(5, 53) = 3.41, p = 0.006, and perceived usefulness, F(5, 53) = 2.51, p = 0.032. Tukey HSD post-hoc tests revealed that starting with Correct/Wrong

Table 15: ANOVA Results for Ease of Use and Usefulness by Demographic Factors

Factor	**Ease of Use**	**Usefulness**
Gender	F(1, 57) = 11.66, p = 2.05e-08	F(1, 57) = 12.21, p = 8.86e-09
School	F(2, 56) = 0.51, p = 0.60	F(2, 56) = 1.82, p = 0.16
Age	F(3, 55) = 0.90, p = 0.44	F(3, 55) = 0.30, p = 0.83
Order	F(5, 53) = 3.41, p = 0.0058	F(5, 53) = 2.51, p = 0.032
Result	F(3, 55) = 0.27, p = 0.85	F(3, 55) = 0.43, p = 0.73

Marks followed by Detailed Feedback and finishing with Correct Answer (order 231) led to significantly higher perceived ease of use compared to starting with Correct Answer, Correct/Wrong Marks, and finishing with Detailed Feedback (order 123), meandiff = 0.47, p = 0.02. Additionally, order 231 was significantly better than starting with Detailed Feedback and then correct answer (order 312), meandiff = 0.54, p = 0.04. Finally order 231 was also better as starting with Correct/Wrong marks and then showing the correct answer, meandiff = 0.70, p = 0.01.

Table 16: Tukey HSD Results for Ease of Use by Order

Group 1	**Group 2**	**Mean Diff**	**p-adj**	**Reject**
123	132	-0.02	1.00	False
123	213	0.23	0.82	False
123	231	-0.47	0.02	True
123	312	0.07	0.99	False
123	321	0.03	1.00	False
132	213	0.25	0.92	False
132	231	-0.45	0.29	False
132	312	0.09	1.00	False
132	321	0.04	1.00	False
213	231	-0.70	0.01	True
213	312	-0.16	0.98	False
213	321	-0.20	0.95	False
231	312	0.54	0.04	True
231	321	0.50	0.11	False
312	321	-0.04	1.00	False

For perceived usefulness, Tukey HSD post-hoc tests indicated that starting with Correct/Wrong Marks followed by Detailed Feedback and finishing with Correct Answer (order 231) was significantly more effective compared to starting with Detailed Feedback (order 312), meandiff = 0.54, p = 0.04.

These results suggest a clear preference for a learning sequence that begins with receiving correct or wrong feedback, followed by detailed feedback, and concluding with seeing the correct answer. Specifically, the sequence 231 (Correct/Wrong Marks, Detailed Feedback, Correct Answer) recorded the highest perceived ease of use and usefulness, highlighting a potentially effective learning loop where students first receive simple feedback, then

Table 17: Tukey HSD Results for Usefulness by Order

Group 1	Group 2	Mean Diff	p-adj	Reject
123	132	-0.103	0.995	False
123	213	0.0997	0.994	False
123	231	-0.3864	0.075	False
123	312	0.1553	0.932	False
123	321	0.0422	0.9999	False
132	213	0.2028	0.958	False
132	231	-0.2833	0.757	False
132	312	0.2583	0.860	False
132	321	0.1452	0.989	False
213	231	-0.4861	0.141	False
213	312	0.0556	0.9998	False
213	321	-0.0575	0.9998	False
231	312	0.5417	0.035	True
231	321	0.4286	0.207	False
312	321	-0.1131	0.994	False

detailed explanations, and finally confirmation of correctness.

An analysis of age-related data shows that perceived usefulness and ease of use do not vary significantly with the age of participants. The ANOVA results for ease of use by age yielded F(3, 55) = 0.90, p = 0.44, indicating no significant differences across age groups. Similarly, the ANOVA results for perceived usefulness by age yielded F(3, 55) = 0.30, p = 0.83, again showing no significant differences. These findings suggest that age does not significantly impact students' perceptions of the ease of use and usefulness of the feedback methods used in this study.

When examining the impact of learning styles on perceived ease of use and usefulness, the differences among the groups are minimal and not statistically significant. The ANOVA results indicate no significant differences in perceived ease of use (F(3, 55) = 0.63, p = 0.60) and perceived usefulness (F(3, 55) = 0.47, p = 0.71) based on learning styles. The Tukey HSD post-hoc tests further confirm these findings, with no significant pairwise comparisons among the different learning styles.

The analysis indicates a significant impact of gender on the perceived usefulness and ease of use of the methods. ANOVA results show F(1, 57) = 11.66, p = 2.05e-08 for ease of use and F(1, 57) = 12.21, p = 8.86e-09 for perceived usefulness, indicating statistically significant differences between the genders in both categories.

Tukey HSD post-hoc tests provide further insights into these differences. For ease of use, significant pairwise comparisons were observed, particularly between participants who chose not to identify their gender and other groups. The mean difference between those who did not want to identify and boys was -2.54 ($p < 0.001$), and between those who did not want to identify and girls was 2.52 ($p < 0.001$). However, it is important to note that the groups "Ape," "Car," and "Does not want to identify" had only one participant each, which can heavily influence the statistical significance and interpretation of these results.

Table 18: Tukey HSD Results for Ease of Use by Gender

Group 1	Group 2	Mean Diff	p-adj	Reject
Ape	Boy	0.38	0.86	False
Ape	Car	-0.17	1.00	False
Ape	Does not want to identify	-2.17	0.0007	True
Ape	Girl	0.35	0.89	False
Boy	Car	-0.54	0.62	False
Boy	Does not want to identify	-2.54	0.00	True
Boy	Girl	-0.02	1.00	False
Car	Does not want to identify	-2.00	0.0022	True
Car	Girl	0.52	0.66	False
Does not want to identify	Girl	2.52	0.00	True

For perceived usefulness, similar significant differences were found. The mean difference between those who did not want to identify and boys was -2.50 ($p < 0.001$), and between those who did not want to identify and girls was 2.46 ($p < 0.001$). Again, the small sample size for these categories should be considered when interpreting the results.

Table 19: Tukey HSD Results for Usefulness by Gender

Group 1	Group 2	Mean Diff	p-adj	Reject
Ape	Boy	0.41	0.79	False
Ape	Car	-0.08	1.00	False
Ape	Does not want to identify	-2.08	0.0006	True
Ape	Girl	0.37	0.85	False
Boy	Car	-0.50	0.66	False
Boy	Does not want to identify	-2.50	0.00	True
Boy	Girl	-0.04	0.99	False
Car	Does not want to identify	-2.00	0.0012	True
Car	Girl	0.46	0.73	False
Does not want to identify	Girl	2.46	0.00	True

Overall, while the statistical analysis shows significant differences based on gender, the limited number of participants in the "Ape," "Car," and "Does not want to identify" categories suggests that these findings should be interpreted with caution. The larger groups of boys and girls, each with around 30 participants, provide more robust data, indicating some nuanced variations in perceptions of ease of use and usefulness across different gender identifications. These genders also had a low difference for ease of use (mean-diff = -0.02, p = 0.99) and usefulness (mean-diff = -0.0414, p = 0.99).

Similarly, the analysis of the school environment shows no significant differences between schools in terms of perceived ease of use and usefulness. The ANOVA results for perceived usefulness by school were F(2, 156) = 0.35, p = 0.71, and for ease of use, F(2, 156) = 0.51, p = 0.60, suggesting that while there may be minor variations, the school environment does not significantly influence the effectiveness of the feedback methods. These findings imply that the methods are generally effective across different educational settings.

The feedback stage of the experiment provided valuable insights into the effectiveness of different feedback methods, despite technical issues such as app crashes and feedback mismatches. Common errors included subscript and superscript problems, capitalization errors, and chemical formula errors, indicating specific areas where students struggled. Analysis of the TAM results showed that perceived ease of use was similar across methods, while perceived usefulness varied. Showing correct or wrong marks and showing the correct answer were perceived as more useful than giving detailed feedback. The order of feedback methods significantly influenced perceived usefulness (F(5, 174) = 2.51, p = 0.032) and ease of use (F(5, 174) = 3.41, p = 0.006), suggesting an optimal learning progression. Additionally, older students reported higher perceived usefulness and ease of use, which aligns with existing research on student feedback dissatisfaction in later years. Learning styles also impacted perceived usefulness, particularly among read/write learners, although no significant differences were found in perceived ease of use or usefulness based on gender (F(1, 158) = 0.09, p = 0.76) or school (F(2, 156) = 0.35, p = 0.71 for usefulness; F(2, 156) = 0.51, p = 0.60 for ease of use). These findings suggest that structured feedback and a gradual increase in feedback detail may enhance student learning and satisfaction.

5.5 Observations

During the course of the experiment, several key observations were made that provided insight into the behavior and preferences of the participants. A notable tendency was for participants to collaborate when they were unsure about what to do. This behavior occasionally led to confusion, especially when the order of experiments varied among the groups. Additionally, it was observed that providing more thorough explanations before starting the experiments significantly improved the participants' understanding of the assignments. Unfortunately, many participants tended to skip the embedded explanations in the app, which often resulted in confusion.

Another observation was the tendency of students to draw or become distracted when they did not know the answer. When they finished their assignments, they frequently engaged in activities they were not supposed to, such as browsing through the tablet. Despite these distractions, participants generally enjoyed the novelty of incorporating new technology into their class activities. The varying responses of teachers also played a role; some were very accommodating and supportive, while others were stricter, insisting that students remain focused.

A significant challenge encountered was the return rate of ethics forms, with only about one-third of participants bringing them back. This low return rate could be attributed to the multiple steps required: taking the form home, getting signatures from the parents and the students, and then remembering to bring it back. Each of these steps presents an opportunity for the form to be forgotten or misplaced.

Overall, these observations highlighted the importance of clear, pre-experiment instructions, the impact of teacher behavior on student focus, and the need for streamlined processes for returning important documents like ethics forms. They also underscored the value of introducing novel elements in the classroom to engage students and the potential benefits of a structured feedback system in educational settings.

6 Conclusion

The research aimed to answer the question, "How can an application be designed to assist high school students in learning to write chemical formulas by hand, effectively integrating input methods, optical character recognition (OCR) processing, and feedback mechanisms?" The study involved 160 students from seven third-year high school classrooms across three different schools, though only 59 participants submitted the necessary consent forms. The participants' interactions, emotions, and engagement were carefully observed and recorded, with data including demographic details, captured images, answers, and responses to TAM and VARK questionnaires.

The input stage of the experiment revealed that the method of input—writing on blank paper, using grid boxes, or directly on a tablet—did not significantly influence the number of correct answers provided by students. However, students showed a clear preference for writing directly on the tablet. Demographic factors such as age, gender, and learning styles did not significantly impact the Technology Acceptance Model (TAM) scores. The exception was the school environment, which had a notable influence on TAM scores, indicating that the environment and resources of certain schools might affect students' perceptions and adaptability to new learning tools. These findings emphasize the consistent effectiveness of different input methods across various student demographics while highlighting the crucial role of the educational environment.

In the processing stage, the development of OCR models encountered several challenges. The initial model, which utilized the MNIST and A-Z datasets, achieved a high accuracy of 97%. However, this model was not directly applicable to chemical formulas due to its focus on recognizing basic characters and numbers rather than complex chemical notation. When attempting to use a model with labels for all possible chemical formula answers, the accuracy dropped to 0.0%. This was primarily due to insufficient data representation and the extremely large and diverse label set, which made it difficult for the model to generalize effectively. A subsequent model aimed to determine whether answers were correct or incorrect, achieving a moderate accuracy of 38%. This model performed well on common or simpler questions but struggled with more complex chemical formulas, highlighting the limitations in handling nuanced chemical notation. The final model, which focused on evaluating individual questions for correctness, also faced difficulties, achieving a lower accuracy of 32%. This low performance was likely due to the need for specific, detailed information to accurately determine the correctness of each answer. These findings underscore the critical importance of having robust and comprehensive datasets that accurately represent the variety and complexity of chemical formulas to improve OCR model performance in this context.

The feedback stage involved evaluating three feedback mechanisms: providing the correct answer, marking mistakes similarly to manual correction, and giving detailed feedback with hints. When participants were shown the correct answer, they were encouraged to reflect on their responses, promoting self-assessment and critical thinking. Common groups of errors included subscript and superscript notation issues, capitalization errors, inaccuracies in chemical formulas, and miscellaneous comments. These reflections helped categorize the types of mistakes students commonly made and allowed them to understand their errors better. TAM analysis revealed that giving detailed feedback scored slightly lower in perceived usefulness compared to providing the correct answer and showing correct or wrong marks. Despite offering structured feedback, detailed feedback was not deemed as useful as the other methods. The perceived ease of use, however, was similar across all methods. The order of experiments significantly influenced perceived usefulness ($F(5, 53) = 2.51$, $p = 0.032$) and ease of use ($F(5, 53) = 3.41$, $p = 0.006$). An incremental feedback approach, starting with no feedback, then detailed feedback, and finally the correct answer, was preferred by participants. This progression enhanced both perceived ease of use and perceived usefulness. Regarding demographics, there were no statistically significant differences found for age, learning style, or school. However, there was a notable difference related to the gender category "prefer not to identify," but this category included only one participant. The differences between the two significant groups, boys and girls, did not show a large difference for either ease of use or usefulness.

The observations from the study revealed several important insights into participant behavior and preferences. It was noted that participants often collaborated when unsure about how to proceed, leading to occasional confusion when the order of experiments varied. Clear and thorough explanations prior to starting the tasks significantly improved participant understanding and performance, as many tended to skip the embedded instructions in the app, leading to confusion. Additionally, when participants were unsure of the answers or had completed their tasks, they often resorted to drawing or engaging in unrelated activities, indicating a need for engaging and structured tasks to maintain focus. The novelty of the app was generally well-received, sparking interest and engagement among students. Teacher responses varied, ranging from highly accommodating to more stringent, affecting student focus and interaction. Lastly, the return rate for ethics forms was low, attributed to multiple stages in the process, suggesting a need for streamlining to ensure higher compliance. These observations underscore the importance of clear instructions, engaging content, and streamlined administrative processes in educational interventions.

Overall, the study concluded that an application designed to assist high school students in learning to write chemical formulas by hand can effectively integrate different input methods, OCR processing, and feedback mechanisms to enhance learning outcomes. Differences between schools were significant, suggesting that the school's environment and resources might influence students' perceptions and adaptability to new learning tools. The OCR system, while facing challenges, showed that models focusing on single character detection performed best, underscoring the need for robust datasets and tailored models. Feedback mechanisms indicated that a progressive approach, starting with simple feedback, then detailed feedback and finishing with the correct answer, was most effective, enhancing both perceived ease of use and usefulness. The observations emphasized the importance of clear instructions, engaging content, and streamlined administrative processes. These findings suggest that a well-designed application, incorporating these elements, can significantly support students in learning to write chemical formulas by hand and improve their overall learning experience.

7 Discussion

This study aimed to develop and evaluate an application to assist high school students in learning to write chemical formulas by hand. The research was divided into three stages: input methods, OCR processing techniques, and feedback mechanisms. The findings provide insights into the preferences and effectiveness of different approaches, as well as the challenges and opportunities in developing educational tools for this purpose.

In the input stage, three methods were tested: writing on plain paper, writing in grid boxes, and writing directly on a tablet. The two paper-based methods were quite similar, and students showed a preference for the tablet method. However, significant app malfunctions limited the completeness of the data for the tablet input method, underscoring the need for a more robust and reliable system. Future work could explore a broader range of input methods and focus on developing a more stable and validated application. Additionally, enhancing the digital interface and investigating other innovative input techniques could further improve student engagement and learning outcomes.

The processing stage involved the development and evaluation of OCR models to recognize handwritten chemical formulas. While the model using the MNIST and A-Z datasets achieved high accuracy, it required additional modifications to handle chemical formulas effectively. This underscores the need for specialized datasets tailored to educational content. The models trained specifically on chemical formulas faced significant challenges due to insufficient data, resulting in lower accuracy. One limitation was that the dataset consisted only of photographed images of chemical formulas. Future work could focus on creating datasets that include both line tracing data from tablet writing and traditional images to improve the models.

Another issue is the difficulty in distinguishing between very similar correct and incorrect answers. Future research should aim to develop OCR systems that can not only recognize handwritten formulas accurately but also differentiate between types of mistakes. This could involve creating models that can classify errors or implementing pre- or post-processing techniques to verify the correctness of an answer. Moreover, determining the correctness of nuanced details, such as a small capital letter next to a large regular letter, remains a challenge. Future efforts should address these complexities by refining the criteria for correct and incorrect answers and enhancing the OCR models to handle such subtleties effectively.

The feedback stage revealed that structured feedback methods and a gradual increase in feedback detail were preferred by participants. This finding aligns with existing research on the importance of feedback in educational settings. However, the study was limited by only implementing three feedback methods. Future work should explore a broader range of feedback types to determine which are most effective. The concept of increasing feedback detail was perceived as useful, suggesting that this approach warrants further investigation.Additionally, randomizing the students' own images for feedback was not effective, as it complicated the process and made the conclusions less reliable. Future research should either develop a complete system that can provide immediate feedback upon completing the first assignment or use standardized questions and answers to create a more controlled test environment. Developing an application that can directly analyze input and provide immediate feedback would simulate a complete interaction and enhance the learning experience. Future research should focus on creating such real-time feedback systems to better support student learning.

Despite the valuable insights gained from this study, several limitations must be addressed. One significant limitation was the incomplete collection of ethics consent forms, which restricted the participant pool. Future research could address this by either involving older participants who can provide their own consent or, more appropriately for this target group, digitizing the consent form process to streamline and simplify submissions. Additionally, the logistics of arranging experiments with entire classes are challenging and time-consuming. Ensuring that experiments can be conducted efficiently with whole classes or in staggered groups is crucial for maximizing data collection within limited time frames.

To enhance the app's effectiveness, integrating social features could make it more interactive and engaging. Incorporating elements that allow students to collaborate, discuss mistakes, and share progress, such as group activities, peer feedback mechanisms, and interactive discussion forums, could significantly improve the learning experience. Additionally, considering that teenagers often engage in playful behavior, ensuring that the setup is robust and can handle such interactions without compromising the experiment's integrity is essential.

By addressing these limitations and focusing on these areas for future research, the application can be further refined to better support high school students in learning to write chemical formulas by hand.

References

Zamfiroiu Alin, Emanuel Herteliu, and Bogdan Vintila. 2012. Human Interaction with mobile Applications. 6 (12 2012), pp. 323.

Antoni Badia, Julio Meneses, and Carles Conde. 2014. Teachers' perceptions of factors affecting the educational use of ICT in technology-rich classrooms. *Electronic Journal of Research in Educational Psychology* 11 (07 2014), pp. 787–808. https://doi.org/10.14204/ejrep.31.13053

Bernard Bahati, Matti Tedre, Uno Fors, and Evode Mukama. 2016. Exploring Feedback Practices in Formative Assesment in Rwandan Higher Education: A Multifaceted Approach is needed. IV IV (01 2016), pp. 1–22. https://doi.org/10.20472/TE.2016.4.2.001

Susan Brookhart. 2008. How to give effective feedback to your students. https://api.semanticscholar.org/CorpusID:59883213

Raduan Che Rose and Jeffrey Lawrence. 2008. Teachers' Readiness to Use Technology in the Classroom: An Empirical Study. *European Journal of Scientific Research* 21 (08 2008).

Fred Davis and Fred Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13 (09 1989), pp. 319–. https://doi.org/10.2307/249008

Neil D Fleming. 1995. I'm different; not dumb Modes of presentation (V.A.R.K.) in the tertiary classroom. *HERDSA* 18 (1995), pp. 308–313. Research and Development in Higher Education, Proceedings of the 1995 Annual Conference of the Higher Education and Research Development Society of Australasia (HERDSA).

Qingke Fu and Gwo-Jen Hwang. 2018. Trends in mobile technology-supported collaborative learning: A systematic review of journal publications from 2007 to 2016. *Computers Education* 119 (04 2018). https://doi.org/10.1016/j.compedu.2018.01.004

Peter Garst, Reeve Ingle, and Yasuhisa Fujii. 2023. OCR Language Models with Custom Vocabularies. (08 2023).

Hodjat. Hamidi and Amir Chavoshi. 2018. Analysis of the Essential Factors for the Adoption of Mobile Learning in Higher Education: A Case Study of Students of the University of Technology. *Telematics and Informatics* 35 (06 2018), pp. 1053–1070. https://doi.org/10.1016/j.tele.2017.09.016

Arthur Hemmer, Jérôme Brachat, Mickaël Coustaty, and Jean-Marc Ogier. 2023. Estimating Post-OCR Denoising Complexity on Numerical Texts. *ArXiv* abs/2307.01020 (2023). https://api.semanticscholar.org/CorpusID:259317132

Brian Irwin, Graham Holden, Stuart Hepplestone, Louise Thorpe, and Helen Parkin. 2011. Using technology to encourage student engagement with feedback: a literature review. *Research in Learning Technology* 19 (12 2011). https://doi.org/10.3402/rlt.v19i2.10347

Kunal Jaiswal, Avichal Suneja, Aman Kumar, Anany Ladha, and Nidhi Mishra. 2023. Preprocessing Low Quality Handwritten Documents for OCR Models. *International Journal For Research In Applied Science And Engineering Technology* (04 2023). https://doi.org/10.22214/ijraset.2023.50664

Srinidhi Karthikeyan, Alba García Seco de Herrera, Faiyaz Doctor, and Asim Mirza. 2021. An OCR Post-Correction Approach Using Deep Learning for Processing Medical Reports. *IEEE Transactions on Circuits and Systems for Video Technology* PP (06 2021), pp. 1–1. https://doi.org/10.1109/TCSVT.2021.3087641

Daniel Lopresti. 2009. Optical character recognition errors and their effects on natural language processing. *IJDAR* 12 (09 2009), pp. 141–151. https://doi.org/10.1145/1390749.1390753

Qingxiong Ma and Liping Liu. 2006. The Technology Acceptance Model. *Journal of Organizational and End User Computing* 16 (07 2006), 59–72. https://doi.org/10.4018/joeuc.2004010104

Mohamad Masrek and Ismail Samadi. 2017. Determinants of Mobile Learning Adoption in Higher Education Setting. *Asian Journal of Scientific Research* 10 (03 2017), pp. 60–69. https://doi.org/10.3923/ajsr.2017.60.69

Jamshed Memon, Maira Sami, and Rizwan Khan. 2019. Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR). (12 2019).

Yakubu Bala Mohammed and Damla Karagozlu. 2021. A Review of Human-Computer Interaction Design Approaches Towards Information Systems Development. *Brain Broad Research in Artificial Intelligence and Neuroscience* 12 (03 2021), pp. 229–250. https://doi.org/10.18662/brain/12.1/180

Jill Naiman, Morgan Cosillo, Peter Williams, and Alyssa Goodman. 2023. *Large Synthetic Data from the arχiv for OCR Post Correction of Historic Scientific Articles*. pp. 265–274. https://doi.org/10.1007/978-3-031-43849-3_23

Esty Nurjanah. 2021. Students' Perceptions about Feedback Practices During Academic Writing Course: A Survey Study. *Jo-ELT (Journal of English Language Teaching)*

Fakultas Pendidikan Bahasa Seni Prodi Pendidikan Bahasa Inggris IKIP 8 (06 2021), pp. 1. https://doi.org/10.33394/jo-elt.v8i1.3397

Everistus Orji, Ali Haydar, ˙Ibrahim Er¸san, and Othmar Mwambe. 2023. Advancing OCR Accuracy in Image-to-LaTeX Conversion—A Critical and Creative Exploration. *Applied Sciences* 13(22) (11 2023), pp. 20. https://doi.org/10.3390/app132212503

Edd Pitt, Margaret Bearman, and Rachelle Esterhazy. 2019. The conundrum of low achievement and feedback for learning. *Assessment Evaluation in Higher Education* 45 (06 2019), pp. 1–12. https://doi.org/10.1080/02602938.2019.1630363

Shruti Rijhwani, Daisy Rosenblum, Michayla King, Antonios Anastasopoulos, and Graham Neubig. 2023. User-Centric Evaluation of OCR Systems for Kwak'wala. *ArXiv* abs/2302.13410 (2023). https://api.semanticscholar.org/CorpusID:257219604

Mario Riojas, Susan Lysecky, and Jerzy Rozenblit. 2012. Educational Technologies for Precollege Engineering Education. *IEEE Transactions on Learning Technologies* 5, 1 (2012), pp. 20–37. https://doi.org/10.1109/TLT.2011.16

Guralnick Robert, Raphael Lafrance, Michael Denslow, Samantha Blickhan, Mark Bouslog, Sean Miller, Jenn Yost, Jason Best, Deborah Paul, Elizabeth Ellwood, Edward Gilbert, and Julie Allen. 2024. Humans in the loop: Community science and machine learning synergies for overcoming herbarium digitization bottlenecks. *Applications in Plant Sciences* 12 (01 2024). https://doi.org/10.1002/aps3.11560

Anna Rowe and Leigh Wood. 2008. Student Perceptions and Preferences for Feedback. *Asian Social Science* 4 (01 2008), pp. 78–88. https://doi.org/10.5539/ass.v4n3p78

Mohamed Sarrab, Ibtisam Shibli, and Nabeela Badursha. 2016. An Empirical Study of Factors Driving the Adoption of Mobile Learning in Omani Higher Education. *The International Review of Research in Open and Distributed Learning* 17 (07 2016). https://doi.org/10.19173/irrodl.v17i4.2614

Katta Spiel, Oliver Haimson, and Danielle Lottridge. 2019. How to Do Better with Gender on Surveys: A Guide for HCI Researchers. *Interactions* 26 (06 2019), 62–65. https://doi.org/10.1145/3338283

Bram Edward Vaessen. 2021. *Students' perceptions of assessment and student learning in higher education courses.* Phd Thesis 1 (Research TU/e / Graduation TU/e). Proefschrift..

Ilie Vali. 2023. The Impact of Technology on Collaborative Learning. pp. 126–141. https://doi.org/10.15405/epes.23045.13

Olivia Wohlfart, Alina Wagner, and Ingo Wagner. 2023. Digital Tools in Secondary

Chemistry Education -Added Value or Modern Gimmicks? *Frontiers in Education* 8 (06 2023). https://doi.org/10.3389/feduc.2023.1197296

Mohammed Yousif. 2024. Enhancing The Accuracy of Image Classification Using Deep Learning and Preprocessing Methods. *Artificial Intelligence Robotics Development Journal* (01 2024). https://doi.org/10.52098/airdj.2023348

Janet Zydney and Zach Warner. 2015. Mobile Apps for Science Learning: Review of Research. *Computers Education* 94 (11 2015). https://doi.org/10.1016/j.compedu.2015.11.001

A TAM Questionnaire Questions

Input stage:

Writing on White Paper

1 I found it pleasant to write my answers on white paper.

2 It was easy to photograph my answers on white paper using the app.

3 I found it convenient to enter my answers into the app by first writing them on white paper.

4 I enjoyed entering my answers into the app after writing them on white paper.

5 I would be willing to use the method of writing on white paper again for future activities with the app.

Writing in Grid Boxes

1 I found it pleasant to write my answers in grid boxes.

2 It was easy to photograph my answers in the grid boxes using the app.

3 I found it convenient to enter my answers into the app by first writing them in the grid boxes.

4 I enjoyed entering my answers into the app after writing them in the grid boxes.

5 I would be willing to use the method of writing in grid boxes again for future activities with the app.

Writing on the Tablet

1 I found it pleasant to write my answers on the tablet.

2 The app responded well to my input via the tablet with a pen.

3 I found it convenient to enter my answers into the app by writing on the tablet.

4 I enjoyed entering my answers into the app by writing on the tablet.

5 I would be willing to use the method of writing on the tablet again for future activities with the app.

Feedback stage:

Receiving the Correct Answer

1 I find receiving the correct answer useful for getting feedback.

2 Receiving the correct answer helps me understand the feedback better.

3 I find receiving the correct answer useful for learning chemical formulas.

4 I find receiving the correct answer effective for learning chemical formulas.

5 Receiving the correct answer is easy to use.

6 Receiving the correct answer is clear.

7 I found the tool for receiving the correct answer easy to master.

8 I find it easy to improve my chemistry skills by using the correct answer.

9 I feel comfortable using the correct answer tool.

Receiving Explanations

1 I find receiving explanations useful for getting feedback.

2 Receiving explanations helps me understand the feedback better.

3 I find receiving explanations useful for learning chemical formulas.

4 I find receiving explanations effective for learning chemical formulas.

5 Receiving explanations is easy to use.

6 Receiving explanations is clear.

7 I found the tool for receiving explanations easy to master.

8 I find it easy to improve my chemistry skills by using the explanations.

9 I feel comfortable using the explanations tool.

Receiving Correct/Incorrect Markings

1 I find receiving correct/incorrect markings useful for getting feedback.

2 Receiving correct/incorrect markings helps me understand the feedback better.

3 I find receiving correct/incorrect markings useful for learning chemical formulas.

4 I find receiving correct/incorrect markings effective for learning chemical formulas.

5 Receiving correct/incorrect markings is easy to use.

6 Receiving correct/incorrect markings is clear.

7 I found the tool for receiving correct/incorrect markings easy to master.

8 I find it easy to improve my chemistry skills by using the correct/incorrect markings.

9 I feel comfortable using the correct/incorrect markings tool.

B OCR model evaluations

B.1 MNIST + A-Z dataset

Class	**Precision**	**Recall**	**F1-score**	**Support**
0	0.62	0.30	0.41	1381
1	0.98	0.98	0.98	1575
2	0.95	0.92	0.94	1398
3	0.99	0.99	0.99	1428
4	0.94	0.96	0.95	1365
5	0.84	0.92	0.88	1263
6	0.97	0.97	0.97	1375
7	0.98	0.99	0.99	1459
8	0.98	0.99	0.98	1365
9	0.99	0.98	0.99	1392
A	1.00	0.99	0.99	2774
B	0.99	0.99	0.99	1734
C	0.99	0.99	0.99	4682
D	0.90	0.98	0.94	2027
E	0.99	0.99	0.99	2288
F	0.97	0.99	0.98	232
G	0.96	0.97	0.96	1152
H	0.98	0.98	0.98	1444
I	0.98	0.99	0.98	224
J	0.98	0.97	0.97	1699
K	0.97	0.99	0.98	1121
L	0.98	0.99	0.98	2317
M	0.99	1.00	0.99	2467
N	0.99	0.99	0.99	3802
O	0.92	0.96	0.94	11565
P	1.00	0.99	0.99	3868
Q	0.96	0.99	0.97	1162
R	0.99	0.99	0.99	2313
S	0.99	0.97	0.98	9684
T	1.00	0.99	0.99	4499
U	0.99	0.99	0.99	5802
V	0.97	0.99	0.98	836
W	0.99	0.99	0.99	2157
X	0.99	0.99	0.99	1254
Y	0.98	0.96	0.97	2172
Z	0.91	0.97	0.94	1215
Accuracy	0.97	0.97	0.97	88491
Macro avg	0.96	0.96	0.96	88491
Weighted avg	0.97	0.97	0.97	88491

B.2 Correct or Wrong per page

Class	**Precision**	**Recall**	**F1-score**	**Support**
Correct1	0.34	0.09	0.14	446
Correct2	0.37	0.04	0.07	413
Correct3	0.00	0.00	0.00	86
Correct4	0.03	0.06	0.04	185
Wrong	0.56	0.79	0.66	883
Accuracy		0.38 (2013)		
Macro avg	0.26	0.20	0.18	2013
Weighted avg	0.40	0.38	0.34	2013

C Statistics tables

C.1 Feedback stage

Table 20: TAM Scores by Demographic, Method, and Order for feedback stage

Demographic	Type	Count	Mean	Std
Age 13	Ease of Use	1	2.27	-
	Usefulness	1	2.17	-
Age 14	Ease of Use	21	2.36	0.67
	Usefulness	21	2.47	0.74
Age 15	Ease of Use	34	2.40	0.62
	Usefulness	34	2.49	0.67
Age 16	Ease of Use	4	2.52	0.39
	Usefulness	4	2.67	0.54
Gender: Ape	Ease of Use	1	2.07	-
	Usefulness	1	2.17	-
Gender: Boy	Ease of Use	27	2.46	0.62
	Usefulness	27	2.59	0.67
Gender: Car	Ease of Use	1	2.00	-
	Usefulness	1	2.00	-
Gender: Girl	Ease of Use	30	2.43	0.46
	Usefulness	30	2.51	0.54
School A	Ease of Use	26	2.33	0.49
	Usefulness	26	2.40	0.53
School B	Ease of Use	17	2.44	0.65
	Usefulness	17	2.59	0.61
School C	Ease of Use	17	2.45	0.77
	Usefulness	17	2.53	0.92
Order 123	Ease of Use	22	2.41	0.42
	Usefulness	22	2.56	0.51
Order 132	Ease of Use	5	2.36	0.54
	Usefulness	5	2.50	0.54
Order 213	Ease of Use	6	2.49	0.54
	Usefulness	6	2.81	0.86
Order 231	Ease of Use	12	2.07	0.98
	Usefulness	12	2.09	0.91
Order 312	Ease of Use	8	2.63	0.35
	Usefulness	8	2.63	0.38
Order 321	Ease of Use	7	2.56	0.65
	Usefulness	7	2.54	0.78
Learning Style: Aural	Ease of Use	23	2.41	0.51
	Usefulness	23	2.52	0.57
Learning Style: Kinesthetic	Ease of Use	24	2.36	0.52
	Usefulness	24	2.51	0.60
Learning Style: Read/Write	Ease of Use	8	2.34	1.02
	Usefulness	8	2.33	1.11
Learning Style: Visual	Ease of Use	1	2.07	-
	Usefulness	1	2.25	-