A VR approach for modelling the assessment bias of primary school PE teachers and educating them about it

Giorgos Hadjidemetriou
s2339994

M.Sc. Interaction Technology Thesis

August 2021

Supervisors:
Dr. ir. Dennis Reidsma
Dr. Katharina Proksch
Prof. Dr. Till Utesch

Interaction Technology
Faculty of Electrical Engineering,
Mathematics and Computer Science

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

UNIVERSITY OF TWENTE. | Human Media Interaction

Abstract

Valid and accurate assessment of primary school children in physical education is important to help them better develop their gross motor skills. However, teachers often are biased using both valid and invalid cues that affect the reliability of their assessment. This thesis project presents the process of creating a Virtual Reality (VR) system to calculate this cue utilization of teachers, model their bias, provide them with feedback about it and train them on how to eliminate it.

This research is based on earlier work in which teachers' biases were modelled based on videos by building upon it and presenting the videos in a more immersive format in VR. It discusses the generation of movements through motion capture, the acquisition of 3D models of children, generating variations of the models and animations and presenting feedback based on the teachers' bias model.

The VR solution is responsible for generating simulations that users will assess, analysing their responses using Brunswik's Lens Model to create a model of their bias and provide them with feedback and training on how well they performed and how to improve their assessment. A user study with 24 participants (female = 13) was conducted at the University of Muenster. It consisted of a usability study of the VR system, the modelling of their bias, an evaluation questionnaire and an interview and discussion of their results.

Our findings suggest that we are headed in the right direction towards modelling the bias and providing feedback to the users as well as training them for it. We created a cost-efficient pipeline that generates a variety of cue combinations for the models and animations with low effort and cost. In addition, our results suggest that the animations show sufficient variation to be used for detecting people's bias. While no statistically significant conclusions were drawn, the interviews suggest that users accept and endorse this technology and approach as an accepted medium, and our system shows promising results as an initial study in this domain. We expect that follow-up studies will yield very interesting results about the added value of VR for this use case and study more about our approach's efficacy and efficiency both long and short term.

Acknowledgements

This project marks the end of a two-year journey. Coming to the Netherlands for a masters degree was a dream come true and especially finishing at the University of Twente with a project that is highly correlated with my interests. The past two years have been an eye-opening experience, giving me much insight into technology and what we are capable of doing, as well as challenging me in many ways to cross my limits and explore what is out there.

I would like to thank from the bottom of my heart my supervisors Dennis and Till for their help, support and guidance throughout and for recruiting the participants of this study. I know sometimes the email exchange might have been a bit excessive but I am glad we had the opportunity to collaborate the past six months and create this project. I would also like to thank Katharina for her time to provide me with feedback and recommendations on how to improve my work. Moreover, I would also like to thank all the participants of my study, who helped immensely with their input. Thanks to Daniel Davison for providing me with all the necessary equipment throughout my thesis project. Finally, I would like to thank my family for their constant support and especially my girlfriend who was there throughout reminding me to take a walk sometimes.

I hope you, as the reader, find this project inspiring and enjoyable to read as much as I did.

Table of contents

Acronyms

VR - Virtual Reality
GMD - Gross Motor Development
PE - Physical education
MoCap - Motion Capture
HMD - Head-mounted Display
I-TECH - Interaction Technology
UI - User Interface
UEQ - User Experience Questionnaire
LM - Lens Model

Chapter 1 - **Introduction**

1.1 Motivation

Teachers' skills in assessment of Physical Education (PE) in primary schools contribute to the capacity to give the right teaching. However, teachers often show a bias in their assessment by employing invalid cues (such as gender, skin colour, cultural background, etc.) and not using valid cues (such as hop, skip, run, etc.) enough. More specifically, we want to benefit primary school children who are still developing their gross motor skills. Through this thesis project, we aim to explore how to uncover the bias of teachers when assessing their primary school students' performance in PE gross motor development exercises. To do so, we are interested in using immersive technologies, such as Virtual Reality, rather than conventional 2D videos or images, to better stimulate their senses. Furthermore, we aim to assist them in better understanding how to better form their judgment by providing them with feedback about their performance.

Primary school education is one of the first steps children take in acquiring knowledge and developing their skills and they should be properly assessed. In addition, they are still developing in terms of their gross motor skills which are important when it comes to boosting their own perceived athletic and scholastic competence (Piek et al., 2006), as well as help them throughout life in dealing with mental, social, emotional, and recreational aspects (Williams & Monsma, 2017). Therefore, it is the educators' responsibility to correctly assess students to determine their weaknesses and strengths. But, biases exist when it comes to grading students, as explored by previous research (Bonefeld & Dickhäuser, 2018; Lekholm & Cliffordson, 2009; Malouff & Thorsteinsson, 2016), which affects the accuracy of the grade that students achieve. These biases are determined by the different cues that teachers employ when assessing their students which are deemed invalid since they have nothing to do with the student's actual performance (Oudman et al., 2018). While there are a variety of invalid cues, in this study we focus on the following: *i)* skin colour; *ii)* gender; *iii)* cultural background; *iv)* economical status; and *v)* disabilities.

A solution to this issue is that teachers need to become assessment literate to be able to correctly apply their assessment knowledge to different aspects of their students' achievements (Stiggins, 1991). As shown in a proof of principle study by Utesch (2020), we can acquire a model of a teacher's bias by applying Brunswik's Lens Model (as discussed by Fiedler, 1996) on their judgment (explained in more detail in Section 2.7.1). To do so, we input the grade a teacher gives to a child and the cues that the child is characterized by (both valid and invalid cues) to the Lens Model (LM). In return, the LM calculation outputs the weights of how much emphasis a teacher is giving to each cue (cue utilization), therefore the resulting LM of the teacher. Utesch (2020) acquired the teachers' assessment for students by showing them videos of students performing Gross Motor Development (GMD) exercises while also providing them with invalid cues related to the students. Through this, he successfully managed to model the teachers' cue utilization and get a clear picture of how their bias is influencing their grading.

However, performing this form of assessment poses different challenges and problems. First and foremost, recording videos of students performing such tasks is an expensive approach with regards to both time and costs (recording equipment, as well as exercise-related equipment, need to be available). Additionally, when having time and cost in mind, generalizing the videos beyond this set of tasks and exercises (PE- and GMD-related), recording them becomes prohibitively expensive (new equipment, additional planning, additional execution time, etc). Furthermore, the videos lack systematic variation of both valid and invalid cues that need to be present to get a better image for someone's cue utilization and therefore better input for the Lens Model. Last but not least, immersiveness is another concern since videos are a simple 2D representation of the real image and teachers do not have the freedom to see a child from multiple perspectives and with similar depth of detail as in a real-life setting.

As an attempt to solve these problems, our study domain of Interaction Technology has much to offer, since we specialize in investigating and improving the interaction between humans and novel technology to solve problems. Therefore, in our opinion, Virtual Reality (VR) is a perfect fit for this scenario. It can provide us with an immersive 3D experience, where users can freely move around in a digital environment, while also having the ability to manipulate what is shown to them in the digital world. For example, we can show 3D models of children performing GMD related tasks, while systematically varying their invalid cue characteristics as mentioned at the beginning of this chapter, therefore tackling the systematic variation issue. On top of that, we can immerse them in a fitting environment with all the necessary environment props needed, and therefore tackling the immersiveness issue. Moreover, with the introduction of fully portable Virtual Reality Head-mounted displays (HMD) devices, the training session can be transferred to the user's setting instead of specialized training facilities. In addition, by using this technology, we will be able to project different execution qualities of GMD exercises on the digital 3D children. The different execution quality animations can be acquired with the use of motion capture (MoCap) of people performing GMD tasks. This gives us more flexibility and freedom as to which exercises are performed by each digital child, have even better control over the variation of the valid cues when combined with the invalid cues, and also focus on specific valid or invalid cues by restricting the variation on the ones we do not wish to test the users on. As for the feedback phase of the system, we can show more detailed visualizations along with text-based feedback, to make the feedback more understandable, creative and easily digestible by the target group. Last but not least, to evaluate this solution we can perform a between-subjects study of a computer environment versus the VR environment. For the computer study, we will use recordings of the VR system so that the users of both groups will be exposed to the same digital children and environment. For evaluating the feedback loop of the VR system, a within-subject study can be conducted for the VR system, where users will be exposed to both a 2D feedback on the computer as well as a 3D feedback on the VR system.

However, applying this solution is not yet possible due to some problems that need to be addressed and tackled first. The first issue is the technical approach of the system (constructive problem). More specifically, the lack of an architecture for the proposed system, meaning how each subcomponent interacts with the other subcomponents, how data are transferred between them and what these data are. Moreover, the more specific details of how to generate the valid and invalid cues of the 3D models and animations and then systematically varying them. Lastly, for the constructive problem, we do not know how to offer an appropriate feedback loop back to

the user to inform them and attempt to educate them about their biases. The second problem that comes up is an empirical one, and more specifically, it is the question of if and how well the system will work in terms of achieving our end goal. For example, are the animations that we portray on the 3D digital children valid? Do experts in the field agree with the feedback that we provide them with (when the system will be realised)? Also, do users feel they can use this system as a training tool and learn from it? And if they do, do they learn from it? Therefore, our objective is to create a system that addresses all the problems that we have mentioned and then evaluate it to understand its value and appreciation from the target group.

1.2 Research Questions

In this chapter, we will present our research questions, through which we aim to understand what the state-of-the-art is, the gaps in our research in terms of design and implementation choices, and explain the final evaluation of our project.

RQ1: What are the underlying theories and models of our project?

a. How is Brunswik's Lens Model used?
b. How is Brunswik's Lens Model applied to teachers?
c. How is Interaction Technology combined with Sports?
d. How can Motion Capture data be varied?

This research question has the purpose to fill the gap between the current state-of-the-art with our project's aim. By understanding how Brunswik's Lens Model is used and how it is applied to teachers, we can find a way to implement it in our project to calculate the bias of teachers. In addition, by learning how Interaction Technology (I-TECH) is combined with Sports, we can uncover other tools that are in the same domain as our project and draw inspiration from them. Finally, with the last subquestion, we aim to understand how to vary the motion captured animations of people, so that we can systematically vary the animations that will be projected on the 3D digital children, and therefore vary the valid cues.

RQ2: What is the reference architecture of such a system?

a. How are its submodules defined?
b. What data need to be collected?

Through this question, we aim to understand the structure of the system, its definition, and understand the contents of, what will be, the black box for the participants. By acquiring this reference architecture, we will be able to produce a better experience for the participants, but also a product that will be easier for us to evaluate.

RQ3: How well perceived and appreciated is the developed system?

a. Can we successfully generate a larger sample of animations by varying the invalid cues and the MoCap data?
b. Is the VR system and the resulting Lens Model sufficient enough to uncover the biases of participants?

c. Is the VR system sufficient enough to trigger a sensible discussion with the participants for their bias and did participants learn from it?

This question aims at evaluating our system and acquiring an initial understanding of whether we are headed in the right direction in terms of creating a system that works. It will not give us a definitive answer if we have successfully managed to educate the participants about their bias, but rather see if the participants accept such a system as an educational tool, and whether they agree with this form of feedback or not. Moreover, we want to investigate if the stimuli we chose to vary are enough to calculate and uncover the biases of our participants and therefore have a sensible discussion with them about it. Finally, we want to see if we can reduce the cost of how many recordings and time are needed to generate a larger sample of animations.

RQ4: How much does the VR feedback phase add to the 2D version of the feedback?

a. Do the visualizations help make the feedback phase more understandable and convincing with the outcome?

This question targets the immersiveness and additional elements that we added to the VR feedback phase when compared to the computer feedback phase. Since we added more interaction and more visual elements, we want to see, through qualitative data, how much more appreciated the additional elements and the resulting system are.

Through the realization of this project and research, we are expecting that we will acquire knowledge and input regarding this problem to help future research regarding this topic. We expect that through our system architecture, design, implementation and study when we apply this solution to a real-life setting, it will show promising results. Furthermore, with the answers to our research questions, we expect that we will discover pieces of information that show signs of success of our approach, while also helping guide future work. In addition, the impact of our work is more long term and outside of our control and will greatly benefit from follow-up research. This additional work, through long-term studies, will show whether teachers actually improve as assessors, therefore give better teaching and personalised attention to children, and finally reinforce children to perform better and improve on their movement skills.

The structure of this report starts by showing the state-of-the-art literature regarding the main axes that drive our study, such as the lens model, existing work on this domain, design patterns, and technologies. Following this, the design of the system will be explained, by showing each module, sub-module and the architecture in general. Then, we will discuss the implementation details, technologies we used and intricacies of the development phase. The next chapter will focus on the evaluation of the system, containing the study we conducted, followed by the results related to the outcomes of the study. Finally, conclusions, discussion of our results, recommendations and future work will be discussed based on the contents of this report.

Chapter 2 - **Background**

This chapter contains Chapter 2 (Related Work) and Chapter 4 (Underlying Theory and Models) from my Research Topics report (Hadjidemetriou, 2021) which was graded as a separate report from the current one.

The problem of inaccurate and/or biased assessment by teachers and the impact that this issue has on children has been extensively investigated in literature. This research, among other things, is mainly about the biases that affect the formative assessment that teachers employ for the topics of physical education (PE) and, more specifically, gross motor development (GMD) when assessing students. Therefore, in this section of the report, we will firstly broadly talk about assessment and factors that relate to it; discuss about practises used and how teachers can become assessment literate in order to provide the best assessment possible to students; then focus down on the biases and possible ways of how to approach and solve these issues; and finally about GMD and PE and their correlation with the other themes.

2.1 Role of assessment

Assessment can have both a positive and negative impact on students depending on how accurate it is. Therefore, it is important to have accurate assessment so that it has as much positive impact as possible. To elaborate on this section we will use the sub-arguments of the importance of good assessment and the importance of inaccurate and/or inappropriate assessment.

2.1.1 Importance of good assessment

Good and accurate assessment can impact children in many ways throughout their school life and the goal of assessment and educators who employ it should always be for the student to benefit from it.

An extensive discussion about this was performed by Fuentealba (2011), where the author discusses the elements to be considered in designing and implementing assessment as well as common challenges encountered during these processes, namely: i) purpose of assessment; ii) domains to be tested; and iii) characteristics of the assessment tools to be employed. Furthermore, the author suggests that by improving the quality of assessment, students can be supported during their learning process and be motivated further by using assessment as a means for guidance and feedback. Finally, the author discusses that by performing a combination of formative (e.g: mockup exams) and summative (e.g: final exams) gives a more complete picture to the teachers on the students' learning level.

Supporting these statements, Bone (1999), states that students can benefit greatly from proper assessment in order to build their self-confidence, provide them with feedback and motivate them. Just by providing feedback to students, they are able to understand their

mistakes and improve further. This can also be achieved by implementing peer assessment and self-assessment in the assessment cycle of a student.

2.1.2 Importance of inaccurate and/or inappropriate assessment

On the other hand, inaccurate assessment can affect students negatively not only while they are in school but also in their future.

Starting with research conducted in Norway for students of ages 13-16 for the course of mathematics, where the findings suggest that when teachers perform easier grading on students' performance, the students' achievements tend to deteriorate. Therefore, high-achieving students are negatively affected by easier grading, whereas this has little-to-no impact on low-achieving students and mainly high-achieving students benefit more from undergoing hard grading. (Bonesrønning, 2004; Bonesrønning, 2008). Additionally, Bonesrønning (2008), showed that boys express more negative responses than girls when undergoing easy grading.

To further expand on this theme, a study by Poorthuis et al. (2015) researching the effect of report grade cards after six months, found that students of both genders expressed lower emotional and behavioural engagement, but also boys who perceived their performance as high, expressed more negative emotions as an effect of their grades.

Following the sub-theme of more long term effects on students of inaccurate and/or inappropriate assessment, a study with people who were born in 1967 in Sweden (Klapp, 2015), reports that graded low-ability students (namely students who have a limit to the learning that is possible for them) who had lower grades through lower secondary school had lower odds to finish upper secondary education compared to low-ability students who were not graded at all. More specifically, graded girls achieved higher grades throughout lower secondary school and had higher odds to finish upper secondary education compared to ungraded girls, as well as graded and ungraded boys.

To understand where the inaccuracies in assessment come from, it is important to have some information on what practices are used in general by educators when assessing students.

2.2 Assessment practises used

In this part of the report, we will briefly discuss the practises employed by different school systems, such as in Norway, Sweden, the US and China with regards to student assessment as well as the possible rubrics that said countries may have.

Starting with Norway (Prøitz, 2013) teachers use national curricula which include guidelines and tests to create local curricula that apply to their school system and the specific subject they want the curriculum for. To accommodate that, the Norwegian government attempts to establish common cultures across schools so that the creation of such curricula is easier and more suitable for each school. Moreover, the results of this research indicate that teachers in mathematics and science find it easier to grade students when compared to other classes since they calculate the points that the students earn by completing sets of exercises.

When comparing Prøitz (2013), to Lekholm & Cliffordson's (2009) work, which is for the Swedish school system, national tests are employed for the courses of Swedish, English, and

mathematics. Furthermore, to better help teachers grade the tests, a national test bank is available with a vast variety of tests and examples of tests and the students' performance.

For tackling the same theme in the US, a study that targeted secondary school teachers and their assessment practises (McMillan, 2001) showed that teachers, to assign grades to students, tend to employ eight factors: i) academic performance; ii) performance compared to a set scale of percentage correct; iii) how much the specific learning objectives were mastered; iv) how much the student tried to learn (student effort); v) ability levels of students; vi) quality of homework completed; vii) the degree to which the student pays attention and participates in class, and viii) use of zeros for incomplete assignments.

Similarly, a study that has as a main focus secondary schools in China (Sun & Cheng, 2014) shows the difference between the assessment approach between the two countries. In this study, teachers mainly take into account six factors: i) student effort and/or attitude towards the class (if for example, a student is high-ability and low-effort, it affects the grade they will receive); ii) what is fair (since educators also have the role of judges); iii) what is beneficial for the student (since educators also have the role of coaches) iv) encouragement; v) improvement; and vi) strictness.

Teachers throughout the world use different kinds of frameworks to assess students, be it using governmental guidelines or local curricula. This is not always the optimal choice since these kinds of assessments may have faults and issues with providing accurate results. Therefore, for teachers to make the correct decisions when it comes to this matter, they need to become assessment literate.

2.3 Assessment Literacy

This theme is strongly connected with what assessment practises are used by teachers and it focuses on how literate the teachers are when it comes to assessing what students know and can do. Thoroughly talking about this matter, Xu & Brown (2016) wrote a literature review paper in an attempt to reconceptualize assessment literacy.

Initially, Xu & Brown (2016), through the work of AFT, NCME, & NEA (1990), list seven factors in which teachers should be skilled at: i) Developing appropriate assessment methods corresponding to instructional decisions; ii) Recognising unethical, illegal, and inappropriate assessment methods and uses of assessment information; iii) Developing accurate and appropriate student grading procedures; iv) Selecting appropriate assessment methods corresponding to instructional decisions; v) Properly understanding and explaining the results of external and teacher-produced assessment methods; vi) Using assessment results to make decisions related to students and the study material; and vii) Share assessment results with involved stakeholders.

Furthermore, they (Xu & Brown, 2016) talk about the findings of using instruments (Plake et al., 1993; Mertler, 2004) used for measuring the knowledge of teachers about these factors and measuring the strength and weaknesses of teachers in Assessment Literacy. In general, it was found that the measured assessment knowledge of teachers was generally insufficient (Maclellan, 2004; Plake et al., 1993; Mertler, 2004).

In their in-depth literature review, Xu & Brown (2016), also talk about how pre-teachers can become assessment literate. This can be achieved by meeting a few conditions, such as

the fact that assessment education should take various forms (DeLuca, 2012; Hill et al., 2014; Mertler, 2009) and that content of assessment literacy training should be subject-area specific so that teachers can choose what to learn (Lam, 2015; Leahy & William, 2012).

As mentioned above, Xu & Brown (2016), mention seven factors that teachers need to be skilled at, although for our research we are mainly interested in three of them which are tightly connected with formative assessment. These three factors are i, ii, and iii. Each of the seven factors, however, could be lacking or problematic for any teacher and therefore they can be formed as biases by teachers in their assessment practices.

2.4 Biases in assessment are an important source of inaccurate assessment

Although teachers follow protocols and predetermined guidelines on how to assess students, biases still exist that affect the judgment of the teacher both negatively and positively.

First and foremost, gender bias seems to be a pressing matter in the way teachers assess their students. Through the longitudinal study of Klapp (2015), they found that there seems to be a bias in favour of female students in all subjects with regards to their grades, hinting at significant gender bias. Another research supporting gender bias is Lekholm & Cliffordson (2009) where the results show that female students perform better in national tests regarding tests on the Swedish and English languages. A possible explanation is that female students perform better in such classes, through personal experience or expectations of the teachers.

Additionally, a study not related to students and teachers but still talks about how people perceive others by their appearance is the work of Greenlees et al. (2005). This issue is still important since teachers may employ such behaviour towards students due to the students' appearance. Through their study, Greenlees et al. (2005), wanted to examine the effect that body language has but also the combination of body language and the type of clothing, on somebody when they are judging their opponent's mental states and readiness in tennis. For the matter of body language only, they found that positive body language is perceived as a more positive mental state. For the second question, they found that when the opponents were wearing tennis sportswear and had positive body language they were perceived more positively than when wearing generic or tennis sportswear and had negative body language. Therefore even the combination of body language and clothing plays an important role in how someone is perceived.

To further expand on the possible biases that exist in grading students, Malouff & Thorsteinsson (2016), through a meta-analysis of 20 studies, where assessors were exposed to specific information about the student, other than their performance (e.g: race/ethnic background, education-related deficiencies, physical unattractiveness and poor quality of prior performance). In this research, the authors concluded that when assessors know irrelevant information about students, then bias can occur, although they did not find the correlation of a specific bias to this kind of attitude.

Bias towards students' grades from teachers can also be detected in more factors rather than gender. An example is ethnicity, where Sprietsma (2009), through a study conducted in Germany, compared identical sets of essays by randomly assigning Turkish or German names

to the essays. Through this experiment, they found that Turkish named tests scored significantly lower grades. More specifically, the study revealed that the more experienced the teacher is, the less good feelings they had towards teaching migrant students.

To further elaborate on this matter, one study that supports this is Bonefeld & Dickhäuser (2018), where it was found that pre-service teachers graded migrant students significantly worse than non-migrant students. They also found that this bias occurs mostly when teachers have to employ less rule-based judgement (e.g: give grades to the students) rather than when they have to employ rule-based judgements (e.g: count the errors of students).

Bias can be an important factor that diminishes the performance of students and it can be visible in many forms, such as body language, appearance, gender, ethnicity, etc. Furthermore, it can be more visible when teachers perform formative assessment, which is also the form of assessment used in subjects related to gross motor development which is the topic of importance in our research.

2.5 Assessment in GMD

One of the aims of this research, apart from showing the importance of the aforementioned themes, is to solve the incorrect and unbalanced assessment of children employed by physical education (PE) teachers. To achieve this, we need to understand what gross motor development in children is and show its importance.

First and foremost, for understanding what gross motor development (GMD) is, Williams & Monsma (2007), define motor development as the ability of children to gradually be able to use large and small muscles of the body and also being able to improve this skill while in adolescence. Through these muscles, children can move their bodies more skillfully and more fluently explore the world around them and interact with objects.

For testing the gross motor skills of children, the Test of Gross Motor Development (TGMD) was proposed in 1985 by Ulrich with two more iterations after that to improve it (Ulrich, 1985; Urlich, 2000; Ulrich, 2017). Through this framework, assessors can test children of ages 3 to 10 to assess their gross motor development regardless if they have any disabilities or not.

To signify the importance of these skills in children, Piek et al. (2006), performed a study on children of ages of 7½ to 11 years old and 12 to 15½ years old. Through their work, they tested children on fine motor skills, which are defined as skills that give you the ability to print, write, perform grooming tasks, etc. (Rose et al., 1997), as well as their gross motor skills. Their results showed that children of these age groups with poor motor skills have issues with their school and athletic abilities, physical appearance and behaviour. More specifically, children with less fine motor skills perceived their school capabilities more negatively and children with less gross motor skills perceived their athletic capabilities more negatively.

2.6 Recent work to tackle these issues

In recent work, Utesch (2020) found that teachers, rather than using only valid cues when assessing GMD development in children (e.g: quality of hop, skip, throw, run, etc.) also employ invalid (biased) cues (e.g: name, sex, clothes, skin colour, etc).

To achieve this, Utesch (2020) used pre-recorded videos of children who performed GMD related tasks which were then assessed by teachers based on GMD related criteria. This

work, although effective, demanded an extensive amount of time to record the videos and their different variations. This leaves room for improvement and innovation to make this process more structured, streamlined and easy.

Moreover, by employing frameworks such as a variation of Brunswik's Lens Model, he managed to identify and model these invalid cues explicitly and made steps to inform teachers about them with the aim to improve their future assessment. After a few iterations, teachers showed clear signs of improvement with regard to their assessment practices.

Through Utesch's (2020) work, it is apparent that teachers can be assisted in recognising these assessment practises and shape them into being more accurate and appropriate for students. The outcome of these improvements and changes in teachers' assessment practises, brings the development of appropriate assessment methods that also correspond to instructional decisions. Utesch (2020), sets the stepping stones for tackling these factors and we believe that with our contribution, we will be able to make the impact even more visible.

2.7 Underlying Theory and Models

While the previous sections discuss more the theoretical background of the concept of this project, in this part we will talk about the different underlying theories and models of what we need to know and understand.

2.7.1 Brunswik's Lens Model

For approaching the problem of solving the bias that PE teachers have towards assessing children and helping them understand their bias, we will use Brunswik's Lens Model framework. In this section, the main principles of the model will be briefly explained and we will also discuss some of the research that has already been made in this domain.

A study that is based on Brunswik's work is proposed by Fiedler (1996). The author discusses that the information that we perceive about someone is not just a singular value but rather a combination (or a vector) of contributing factors which also include errors and noise. These errors and noise are the bias in the way someone perceives someone else. Moreover, Fiedler (1996) concluded that the more observations we have for someone, the more these factors are observed. Therefore, the bigger the sample size of observations, the more the effect of bias is visible in someone's perception of someone else.

To better understand the construction of the Lens Model, Douglas & Gifford (2001) provide a brief explanation of the traditional model. The model has on the one side, the ecological validity, which holds the true measures of the environment and how the environment is, and on the other side, we can find the cue utilization which is the cues that the participants use to rate their perception of something, and their weight. Finally, the cues that can be used to describe the environment are present in the centre of the model, with both the ecological validity and the cue utilization sides connected with the cues on each side. The connections can also show the weight of the connection (utilization of the cue).

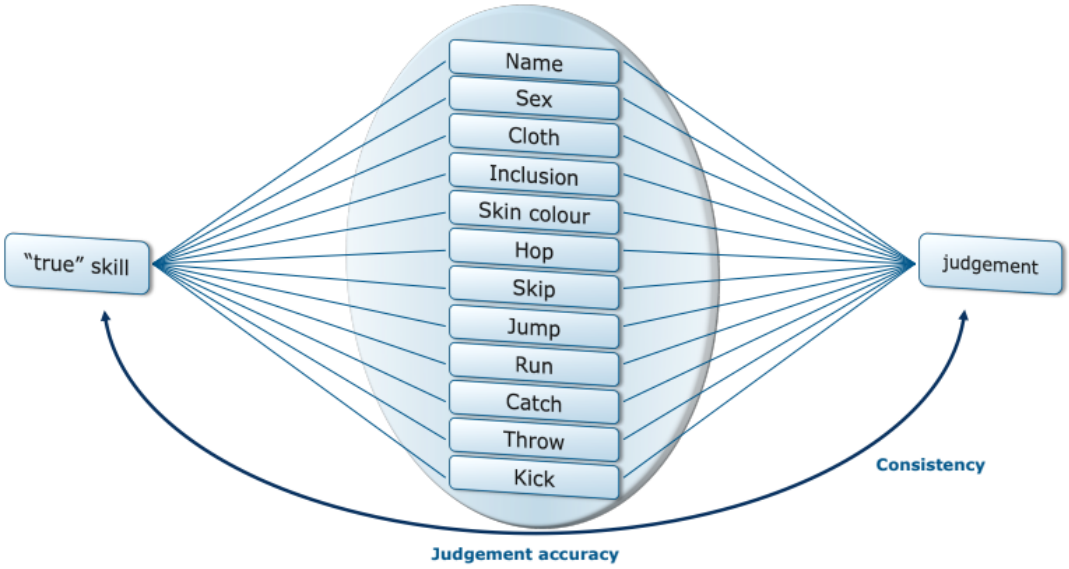


Figure 2.1. The Lens Model for assessing primary school children. Adapted from Utesch (2020)

Figure 2.1 demonstrates the Lens Model when assessing primary school children in GMD related exercises. In the centre of the model, we have the cues that characterize a child. On the left side of the model, we have the "true" skill of the child (distal variable), which is the exact performance or value of the child's characteristic regarding the specific cue. On the right side, we have the judgment of the assessor and how they perceive the child's cues (subject's judgment). The result of the Lens Model is the weights (coefficients) of the assessor's perception of the child's performance (the weights of the lines connecting the judgment of the assessor and the cues).

The Lens Model can be quantified and calculated through a Lens Model Equation which is multiple regression analysis. One of the first Lens Model Equations was by Hammond et al. (1964) and is the following:

$$r_a = \frac{R_e^2 + R_s^2 - \Sigma d}{2} + C\sqrt{(1 - R_e^2)(1 - R_s^2)},\ where\ \Sigma d = \Sigma(r_{e_i} - r_{s_i})(\beta_{e_i} - \beta_{s_i})$$

r_{e_i} is the correlation between the distal variable and the subject's judgment.

R_e is the multiple correlation for the environment side (left side) of the lens model, which is the multiple correlation between the distal variable and the cues.

R_s is the multiple correlation for the subject side (right side) of the lens model, which is the multiple correlation between the subject's judgment and the cues.

r_{e_i} is the correlation between Cue *i* and the distal variable.

r_{s_i} is the correlation between Cue *i* and the subject's judgment.

β_{e_i} is the standardized regression weight for Cue *i* from the regression of the distal variable on the cues.

β_{s_i} is the standardized regression weight for Cue *i* from the regression of the subject's judgment on the cues.
$C$is the correlation between the residuals from the two regression equations.

This theoretical model can be used in a variety of areas in order to understand people's perceptions of specific factors. An example of this theoretical model is Douglas & Gifford (2001), where students and professors rated different classrooms in terms of friendliness and overall preference. This was achieved by using seven cues to describe the perception of the classroom environment by the participants. With this research (Douglas & Gifford, 2001), we can conclude that the Lens Model can be used not only for acquiring the perception of people for other people, but also other entities as well.

Furthermore, Koch's (2004) aim was to investigate the construction process of people's perception of the gender of others and understand the cues that lead to these conclusions. For this experiment, they used pragmatic, semantic and syntactic cues and each one of these categories had a variety of sub-cues. Through this research, we can see the link between people's perception and the cues they use to assign traits to unknown individuals and how this use of cues varies among genders of both sender (people in the experiment) and the recipient (participants of the experiment).

For our research, with the use of this model, we expect that we can better understand the cue utilization of teachers when it comes to grading students, to train teachers on how much each cue is used and which cues are used more specifically, therefore eliminating the bias and training teachers as better assessors. To discuss steps that have already been taken towards this direction, the following subsections will talk about some already existing works.

2.7.2 Brunswik's Lens Model applied to teachers

To understand teachers' behaviour towards children based on factors that are available to said teachers, Cooksey & Freebody (1987) wanted to investigate the contribution of cues to the judgment of kindergarten children reading skills by teachers. To achieve this, they used 5 cues that spanned across two main categories: i) Demographic factors; and ii) Child's cognitive abilities. Cooksey & Freebody (1987), came to the conclusion that teachers' judgments have an impact on how they behave in class. So, if for example, a teacher predicts a potentially good student as a low achieving one, then the teacher's behaviour will adapt to treat the student as low achieving. Furthermore, they found that the cognitive ability cues had more impact on how teachers form their judgment. The authors also expected that if the information about the cues and how teachers form their judgment was made available to teachers, then this would improve their judgment accuracy on forming initial expectations of students.

Steps towards this direction were also taken by Oudman et al. (2018), where they aimed to identify the effects of cues such as the name of students, the answers to practise assignments and the combination of the two, to the perceived understanding of students of a specific mathematical concept, by their teachers. Through their results, it is apparent that by only knowing the name of a student, it is enough for teachers to rate the understanding of their students on a concept. This work further implies that more simple cues, such as the name of a student, can also affect the opinion of a teacher for that student.

For our research, we are interested in GMD related teaching and how teachers grade their students. For this domain, Utesch (2020) through extensive research with videos of children performing such tasks, has identified the cues that teachers are utilizing to perform their assessment and how much weight they associate with each cue. The final set of cues included not only cues that describe the overall performance of the child (valid cues) such as hop, skip, jump, etc., but also invalid cues that are irrelevant concerning the child's performance (invalid cues), such as name, sex, clothing, etc. Furthermore, he also examined the case where teachers were provided with feedback about their assessment and how they utilized invalid cues among the valid cues. This was a fruitful attempt since teachers were able to improve their grading to not utilize invalid cues at such a high level.

Concluding, the aforementioned researches are promising in terms of what we are aiming to achieve through our work, with clear examples of being able to identify the cues that teachers use but also the ability to assist teachers in correcting their assessment skills.

2.7.3 Sports I-TECH

Combining Interaction Technology (I-TECH) with sports has been extensively researched previously in the literature (Mueller & Young, 2018; Reilly et al., 2009). Many examples of such technologies are already active in the sports domain, such as the Hawkeye system (Baodong, 2014), the Cyclops system for tennis (Chi, 2008) and the Bowling foul-line detector (Chi, 2008).

For our research, we are interested in combining I-TECH with PE and provide teachers with a tool to train in that domain that also provides them with feedback on how to further improve their assessment skills. Our focus can be broken down into two subdomains related to sports and PE more specifically: i) combining I-TECH with PE; and ii) provide a tool for PE teachers to train with for better assessing student's performance.

For the first subdomain that occurs research work has been done by Preuschl et al. (2010), where they developed a mobile system that receives inputs from the way a student performs an exercise and gives feedback to the student based on their physical effort in order to help motivate them. Furthermore, Preuschl et al. (2016), based on the mobile system of Preuschl et al. (2010), proposed methods for fair grading of students based on their physical activity during PE, therefore using the tool for helping teachers acquire better knowledge about their students' performance.

As for the second subdomain of our focus, the application of I-TECH on PE to create a continuous feedback system from the student's performance to the teacher to better educate the teacher is still a challenge that has not been tackled, even though it is a recommended practise to get students to be more engaged with physical education (Castelli et al., 2015). Cushion & Townsend (2019), through their extensive literature review, came to the conclusion that such systems need further development and longitudinal research in order to determine their effectiveness and the impact they have on coaches' (in our case PE teachers) learning. Utesch (2020), through his work, has shown that even using videos as a form of understanding the way teachers assess students is enough to determine the bias that is present, but this is a time demanding and lacking variation in the possible student scenarios.

Therefore, there is a clear gap in the literature as well as practise on how to better establish the feedback loop between teachers and students of PE to provide better training to teachers on how to become more assessment literate.

2.7.4 Varying Motion Capture (MoCap) Data

In order to acquire a vast amount of accurate body movements that teachers will assess, we can record actual people using motion capture. To further enhance it and reduce the amount of motion captured body movements needed, we can apply variation algorithms that can generate an abundance of variations of a model's movements.

First and foremost, Boukhayma & Boyer (2017) focused on dynamic human shapes and how they can generate variations of captured 4D models automatically. 4D models refer to dynamic models which also have movement (kinematic) data. To achieve this, Boukhayma & Boyer (2017) used Gaussian Process Dynamic Models to build a probabilistic dataset of poses which they then used to synthesize an unlimited number of variations. Using this method they also managed to blend versions of the input movements and dynamically generate even more complex poses.

Further work towards dynamically blending frames to achieve a desired final sequence was done by Pan et al (2010), where they generated human-like motions in environments with tight spaces which were filled with obstacles. This was achieved with the help of a decomposition planner that was used to compute the path of low-DOF components of the animation and then generate a trajectory for the animation that would also satisfy specific kinematic and dynamic constraints of the model that was used. Finally, by using a trajectory refinement method, they managed to refine the generated motion from the decomposition planner with mocap data, and therefore make the animation smooth and realistic. This is a promising work in how to dynamically combine animations with specific constraints in complex environments to achieve realistic animations.

Similarly to Pan et al (2010), Boukhayma & Boyer (2018), proposed a method to combine animations in order to end up with a smooth desired animation, while also saving data about the appearance of the shape of the model in a compact form. For our research, we are mainly interested in the first part of their method. To achieve this, they used the shape pose distance to calculate the best next pose to use. After that, by using shape pose interpolation, they performed the interpolation between the initial and the computed pose, and finally using a shape motion transition, they generated a smooth interpolated transition between the two poses.

For the matter of the appearance of the digital models, variation can be achieved by applying the same 3D motion-captured movements onto different pre-existing textured 3D models, such as the ones found at Mixamo. It is important to vary the appearance as well since multiple factors of digital agents influence the perception of people who have to interact with them. Such factors include gender (Shiban et al., 2015; Baylor & Kim, 2004), ethnicity (Baylor & Kim, 2004), hairstyle, clothing and age (Baylor & Plant, 2005; Rosenberg-Kima et al., 2008; Plant et al., 2009).

We can draw inspiration from the researches that have been discussed in this section and through the details the authors propose, acquire knowledge on how we can apply their frameworks to our case scenario and generate a multitude of poses dynamically and on the spot. This can greatly increase the length of the dataset of models that can be used to show to teachers in order to assess their assessment behaviour.

2.8 Summary

In this chapter, we saw the role of well backed and accurate assessment and how it affects children in their early stages of learning as well as further down in their lives. Furthermore, we highlighted different assessment methods that are employed by teachers in different parts of the world, in a variety of subjects. We emphasised what assessment literacy is and how teachers can become assessment literate and then turned our focus towards the different biases that influence assessment. Following this, we discussed how assessment is performed in GMD related tasks and recent work that attempted to set the initial groundwork for solving the inaccurate assessment issues in GMD assessment.

This background section would not be complete without investigating the theoretical and practical frameworks behind the main axes of our research. We focused on Brunswik's Lens Model which is one of the main drive forces of our work, while also seeing how it is applied to teachers. We investigated the combination of Sports and Interaction Technology so we can understand how similar research was conducted, and finally, we discussed acquiring and varying motion capture data to systematically vary the animations that we will record.

The next section will target the design phase of our system, which includes taking the background theory and incorporating it into our design.

Chapter 3 - **Design**

This chapter will focus on the design choices we had to make during the initial phases of the system, but also throughout the later phases to adjust the system. This chapter will discuss in a total of six subchapters how our architecture was designed, what are the software and hardware that we used, what framework we used for the GMD exercises and how we used MoCap to capture the animations, what is the story behind the children models that we present, how the environment and User Interface (UI) are structured and finally, what is the design of the computer control version.

3.1 Architecture

The first steps before realizing our solution were to think of the different components that need to be created and what type of communication happens between them. This led to creating a system architecture diagram (Appendix A). This architecture went under multiple iterations of redesigning until its final shape which is presented in this report.

The core of the system consists of four main subsystems (as seen in Figure 3.1) namely: *i)* Test Battery system; *ii)* Analysis system; *iii)* Feedback system; and *iv)* Training system. In this subsection, we will explain the design choices of what each subsystem is tasked to do.

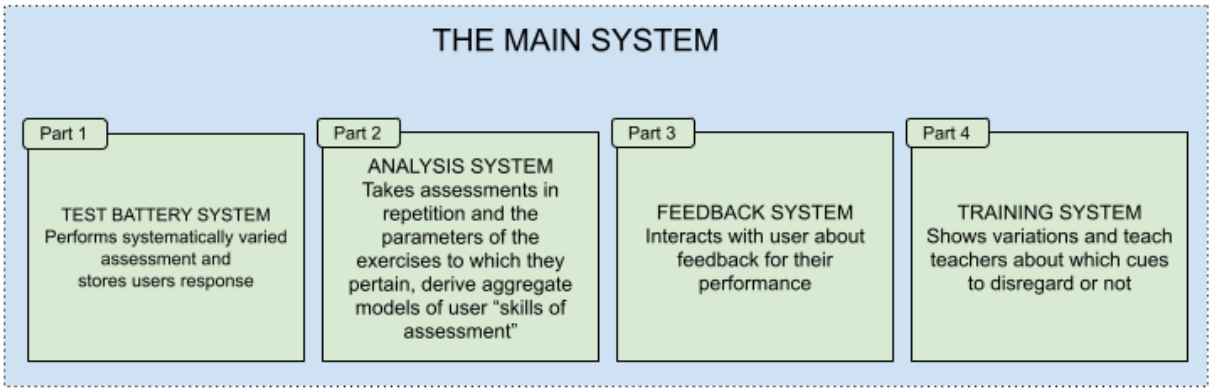


Figure 3.1. Abstract system architecture

3.1.1 Test Battery System

Our aim with this subsystem is to generate the animations and models that the user will have to assess using systematic variation and to store the results of this assessment so that the next subsystem can take over. More specifically, this subsystem consists of four main modules as seen in the grey boxes and two main data storages marked as blue cylindrical containers in Figure 3.2. Starting from the *Simulator of movement with a set of valid/invalid cues*, which is the main module of this subsystem. It is responsible for applying the animations that are generated by the *Animation System* on the 3D human models that are retrieved from the 3D human model library. This library includes the combination of rigged models and the cues that they represent and they are run through the *3D model variation algorithm* to vary the invalid cues on the

models that will be displayed to the users. As for the *Animation System* module, it is responsible for requesting specific exercise animations varied on specific cues from the *Mocap variation algorithm* module. Furthermore, it provides a library of animations of all the exercises to the Simulator. Lastly, the *Mocap variation algorithm* module is responsible for getting the motion capture animations from the data storage, and parametrizing them based on valid and invalid cues, such as random noise or smoothness. We made a rough implementation of a simple variation that will be explained later in Section 4.1.1.

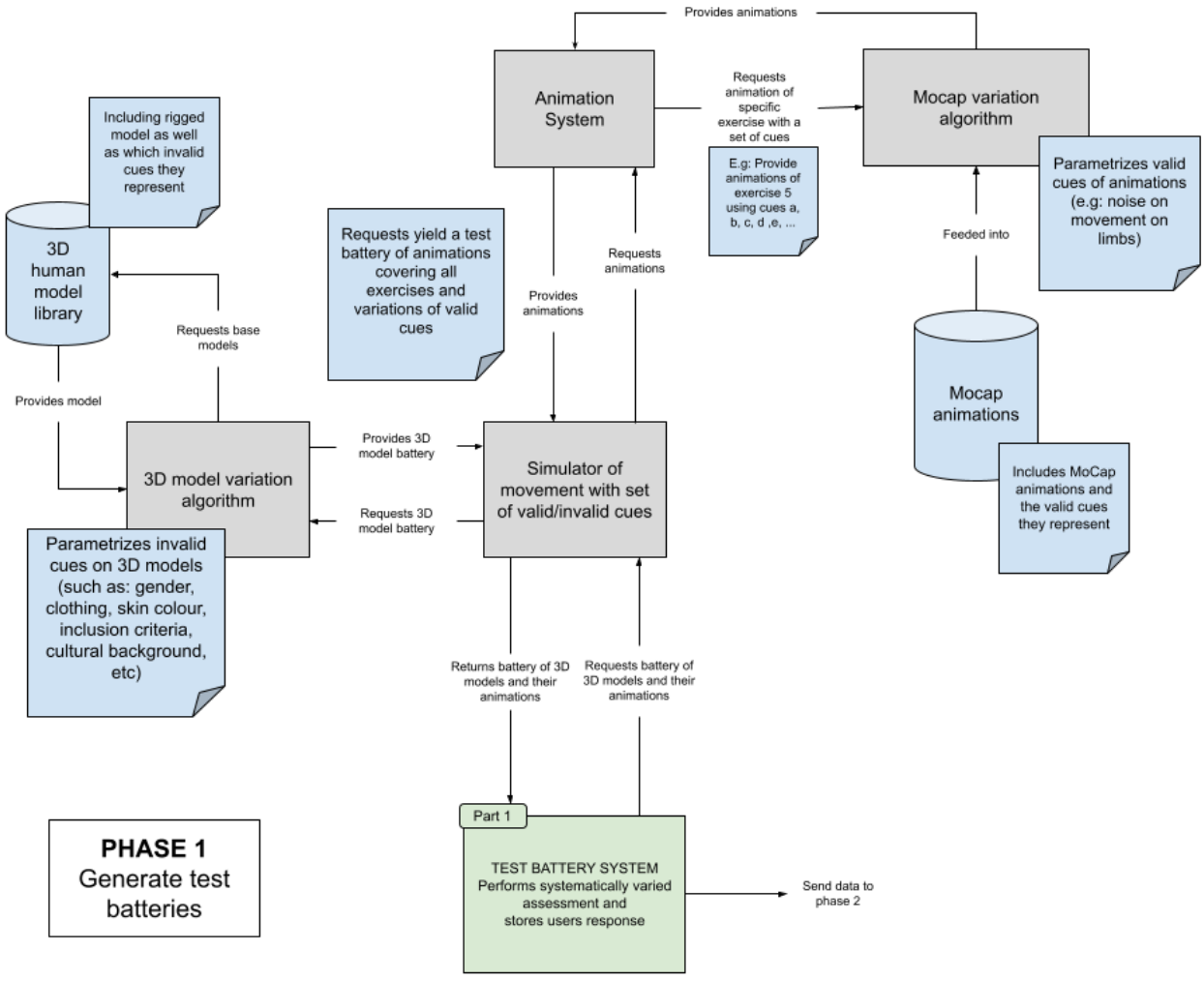


Figure 3.2. Part 1 of the system

3.1.2 Analysis System

The second subsystem of our architecture is the *Analysis system* (Figure 3.3). This module's responsibility is to take the assessments of the users as input, after they evaluate the previously generated 3D models, and derive models of the users' bias based on the Lens Model (details in Section 4.1.2). The bias models of the users (combination of valid and invalid cues, the user's assessment and the resulting Lens Model) are then stored in a data storage so that they can be used later. For the Lens Model implementation, we needed in total a minimum of 60

inputs from the users, meaning that they had to assess 12 different children, and each child would perform a set of five exercises. On each different child, we varied the invalid cues, and on each exercise variation, we varied the valid cues. This process will be further elaborated in the next chapter.

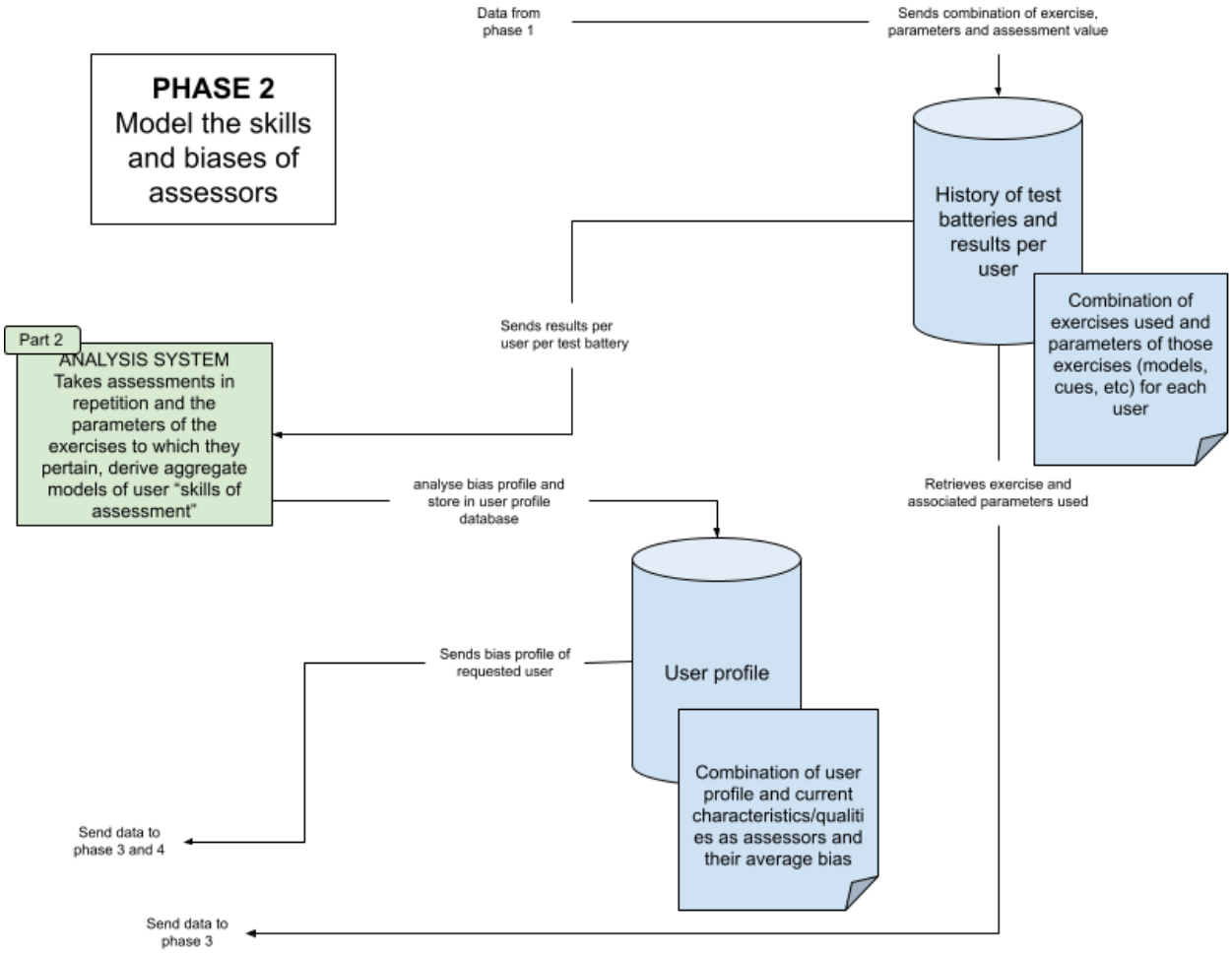


Figure 3.3. Part 2 of the system

3.1.3 Feedback System

The Feedback System (Figure 3.4) is responsible for showing the users a model of their bias. This phase is something that is purely on the front-end of the system, with no intricate back-end functionalities. It is considered as a separate module though since it is an important individual component of the system. More details follow in Section 4.1.3.

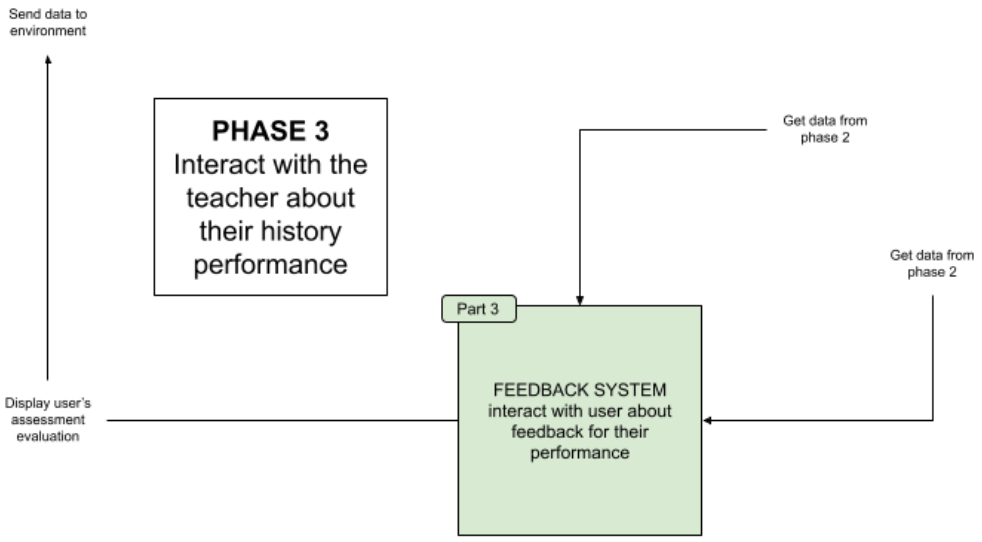


Figure 3.4. Part 3 of the system

3.1.4 Training System

Lastly, the *Training System* (Figure 3.5) is responsible for showing variations of the models that the users assessed and provide insights on what they should pay more or less attention to improve their assessment. This submodule requires some processing with regards to generating the variations of the models that will be shown. More details follow in Section 4.1.4.

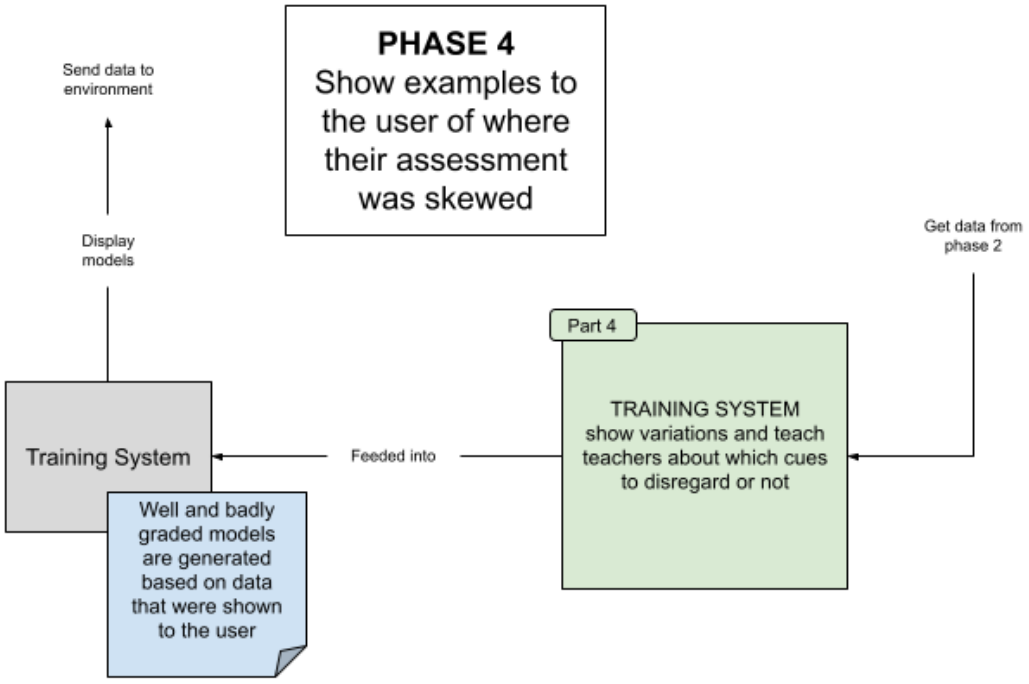


Figure 3.5. Part 4 of the system

3.2 Software & Hardware

To go into further detail regarding the design of our system, we first need to talk about the technologies that we decided to use since they will provide background on why we made specific decisions regarding the rest of this chapter.

Starting with the hardware side of things, since we are interested in developing a VR solution, we had to decide which one of the available HMDs we would use. The choice was quite straightforward and we settled on using the *Oculus Quest 2*, since it offers full mobility, without the need of having additional hardware equipment to run the simulation. The only requirement is a stable internet connection. This opened the possibility of performing user studies at the location of the users, instead of having to bring users to the lab, which was also something we wanted to avoid due to COVID-19. The other hardware requirement was the motion capture equipment. The facilities at the HMI lab offer a medium-sized volume capture area setup that uses the *OptiTrack Prime 13*[1] series, which is directly connected to a desktop computer in the lab. We, therefore, used that since it was readily available.

Moving on to the software choices, we had more flexibility as to which software we would like to use. Initially, we had to choose which game engine to use in which we would develop the system. The options here were *Unreal Engine*[2] or *Unity*[3]. We chose to work with Unity since it is the game engine that we are more familiar with. Secondly, the programming languages that we used was *C#* which is natively supported by Unity, as well *Typescript* and *R*, which were used for creating the server of our system and finally *Angular* for creating the computer version. Thirdly, we had to decide on the 3D modelling software that we would use for creating the assets that would be presented throughout the simulation. For this, we decided to use *Blender*[4], since it's open-source, with large community support. Lastly, the software responsible for cleaning up and relabelling the animations that we would retrieve from the MoCap sessions is *Motive:Body*[5]. We did not have a choice for this software since it is the one shipped with the specific hardware that we used for the motion capture.

3.3 GMD Framework & Motion capture

For acquiring the motions that the digital children would perform in the system, we had to decide first on the GMD framework that we would use. We found frameworks from a variety of organizations such as *kenniscentrum*[6], *mobak*[7] and *hsrpsychology*[8]. The drawback was that these and many others were either for educational purposes only or bundles that you need to pay for. Therefore we decided to work with the *TGMD-3* (Ulrich, 2017) which targets children aged 3-11 and tests for both locomotor and ball throwing skills and in the end it evaluates the overall gross motor skills of children. Furthermore, this test was used by Utesch (2020) for evaluating their method which is the main inspiration for our work.

[1] https://optitrack.com/cameras/prime-13/
[2] https://www.unrealengine.com/en-US/
[3] https://unity.com/
[4] https://www.blender.org/
[5] https://optitrack.com/software/motive/
[6] https://tools.kenniscentrumsportenbewegen.nl/sportfolio-internationaal/onderwerp/measuring-childrens-motor-skills-one-minute/
[7] http://mobak.info/en/mobak/
[8] https://www.hsrpsychology.co.uk/services/specific-assessments/physical-and-sensory/motor-skills-assessment/

The TGMD-3 offers a plethora of exercises such as *run, gallop, hop,* etc for the locomotor skills and *two-hand strike of a stationary ball, one-hand stationary dribble*, etc for the ball throwing skills. Due to limitations to our motion capture space at the lab, we had to restrict our exercises to the ball throwing ones only, since they are executed while standing in place (Figure 3.6). Furthermore, from the subset of the ball throwing exercises, we removed the *two-hand strike of a stationary ball*, since it is an exercise mostly present in US schools (due to the exercise's baseball nature) and not representative of the European school system. Furthermore, we had to remove the exercise *kick of a stationary ball* since it was the only one that required movement.

Figure 3.6. Performing MoCap at the HMI lab with one of the actors

3.4 Children Models

The digital representation of children (Figure 3.7) is the most important aspect of our system since it is what the users would need to assess in order to successfully model their bias. We wanted to avoid 3D modelling our own models of children, due to lack of experience and limited time. Therefore, we found the *o3n UMA Races Stunner Jane & John Standard Render Pipeline*[9] asset on the Unity asset store, which allows for full customization, in terms of age, height, body weight, etc of humanoid 3D models.

As for the elements of the digital children that we could not control over the asset, such as clothing and skin colour, we had to fabricate our own. We wanted to introduce variety to both to achieve multiple combinations in the end. The process for how these were created will be explained in the next chapter (Implementation, Section 4.3).

[9]https://assetstore.unity.com/packages/3d/characters/o3n-uma-races-stunner-jane-john-standard-render-pipeline-175172

Figure 3.7. One of our digital children models with custom clothing, based on the o3n UMA model

3.5 Environment & UI

Virtual Reality has the additional benefit of immersing the user in a full-fledged virtual environment. We wanted to capitalize on this aspect and place our users in a familiar and fitting environment for what we wanted the simulation to be. Therefore we decided to use an indoor basketball/sports hall as our environment (Figure 3.8). Since we wanted to test on PE exercises, such an environment is fitting, as it resembles the environment of actual school sports halls.

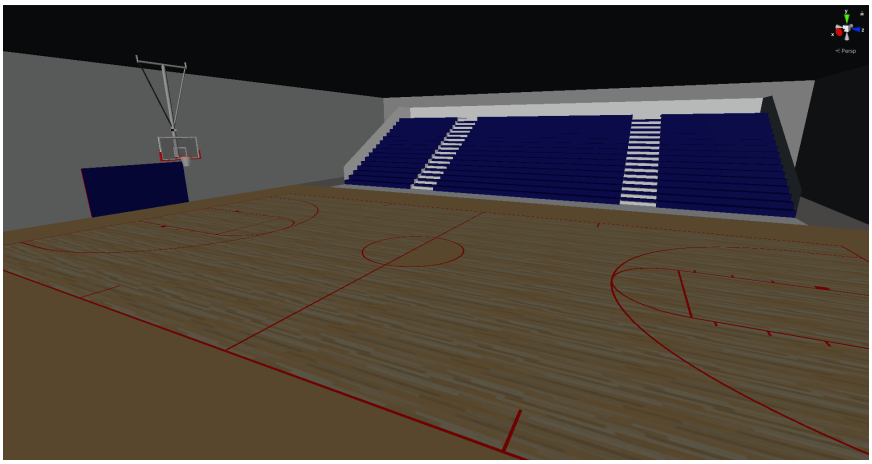


Figure 3.8. The basketball hall environment of our simulation

Furthermore, in the simulation, we wanted to supply more details to the users regarding the characteristics of the digital children, such as their name, exercise they are performing, whether they have any inclusion criteria (NL: kinderen met rugzak, DE: inklusionkinder) and gender. To tackle this, we decided to create a UI element that would hold these details for the users to read on demand (Figure 3.9). This will be explained in detail in Section 4.4.

Furthermore, we wanted to give options to the users as to how they can control the animations that the digital children are performing, such as pausing the animation. This would allow the users to see the child in a specific pose, as well as rotating the child around and skipping to the next or previous movement exercise of the child (Figure 3.10). For that, we

added a media key panel on the left controller of the user, which is a feature that is vastly used in VR games and experiences.

Lastly, for the users to input their assessment for the digital children, we added an additional UI panel on the left controller, over the media keys panel, through which users can do so (Figure 3.10). This panel also includes the count of which child simulation the user is currently assessing, as well as which exercise the child is performing. They can also save the grade they want to submit, as well as submit the assessments of the child to move to the next child.

Figure 3.9. One of our digital children models with the characteristics panel

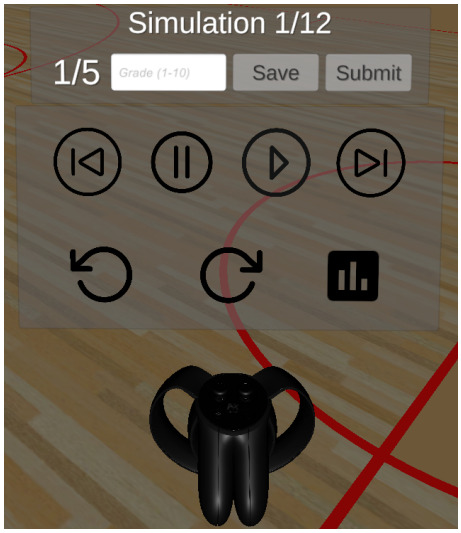


Figure 3.10. The media key and assessment information panels on the left wrist

3.6 Computer Control Version

As for the computer side of things, the control version's design was straightforward (Figure 3.11). We created a website that contains a video player that contains four angles of the digital children who are performing the exercises, along with an information and instructions

panel on the side of the video player. At the bottom of the page, we added navigation buttons for navigating through the exercises and also the input field for the assessment grade as well as the save and submit buttons for storing and submitting the assessments. At the top of the page, we added indicators for the simulation and exercise count. The design is simple, yet really similar to the VR version in terms of what content we show to the users.

Furthermore, we had to include a feedback phase UI for the control version to compare the experience of the feedback between VR and the computer. For this, we created a bar chart screen (Figure 3.12) where users can see generalized messages of their feedback in the same fashion as the VR feedback phase (more details in Section 4.1.3).

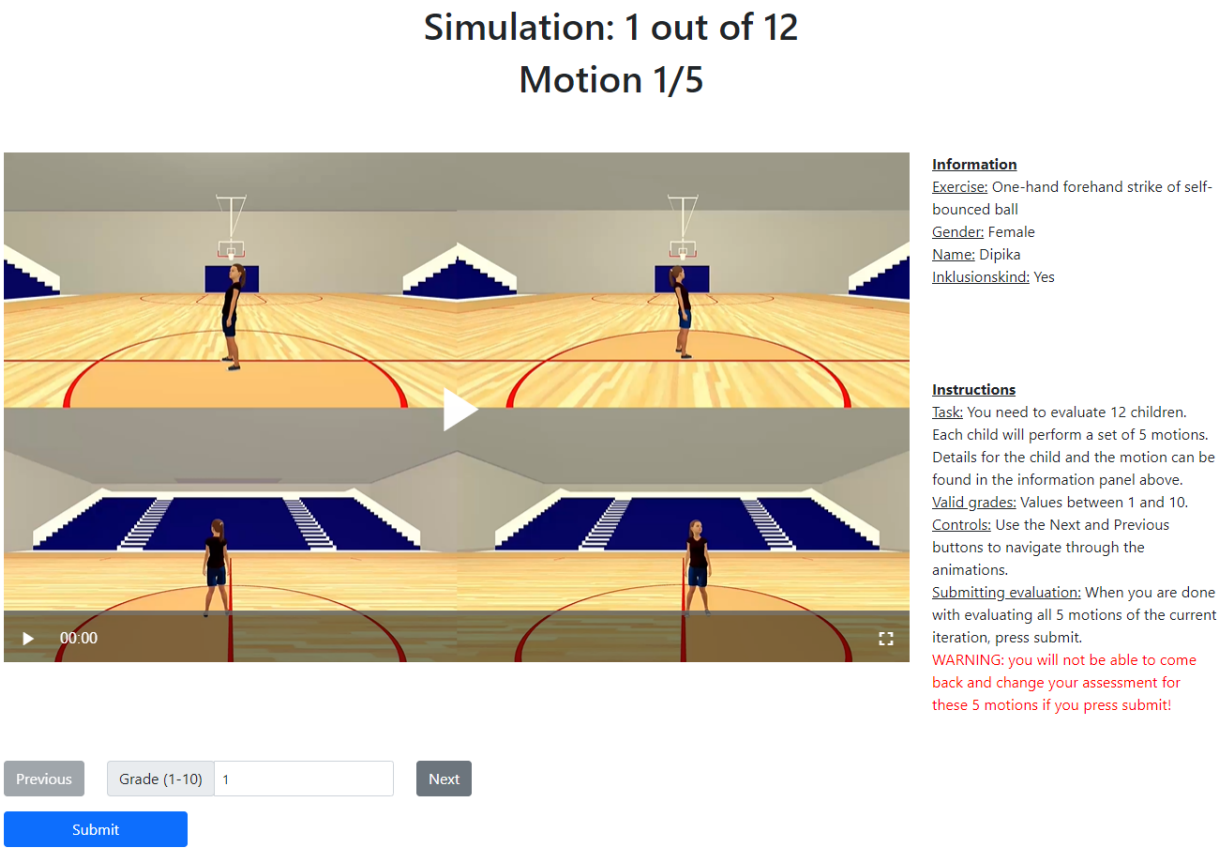


Figure 3.11. The evaluation phase of the computer control version of our study

Figure 3.12 Feedback and training phase of the computer control version of our study

Chapter 4 - **Implementation**

In this chapter, we will explain the processes we followed after our design choices that were discussed in the previous chapter to realise our system and the technicalities behind our approach. It is separated into six subchapters that focus on implementing the architecture of our system, along with acquiring the motion capture, how we created the digital children models, how the environment and the UI were created, how the computer control version is structured and finally how our back-end server was created.

4.1 Architecture

4.1.1 Test Battery System

Starting off the section of the architecture, we will discuss the generation of the battery of simulations that the users will assess. We had to find a systematic way to vary the animations in both appearance and movement so that we have control over the variation that the users were seeing in the simulation.

First and foremost, we have to think of the invalid cues that we would test our users on which are: *i)* gender; *ii)* cultural background; *iii)* skin colour; *iv)* inclusion criteria; and *v)* clothing. Starting from the cultural names, we generated a list of 20 local cultural names (ten female, ten male) for our target group (Germany) and a foreign cultural name list (ten female, ten male) from countries specifically in Eastern countries. The reason is that they are the most distinctly different names from European and Western names. Each subset was added to a list, for a total of four lists, and each one was randomized. We then selected the first three names out of each list (since we would have 12 children in total) to get the 12 different names we would need.

Moving on to the gender and inclusion criteria variation process, which followed the same process for both. For the genders, we have two genders and 12 children. Therefore we created a list with alternating a female and a male child, therefore six female and six male, and then randomly sorted the list. As for the inclusion criteria, since a child either has or does not have inclusion criteria, we generated a list with alternating Yes/No values and followed the same process as the gender list.

The next cue we will discuss is skin colour. We had four different skin colours (explained in more detail in Section 4.3), but we wanted to have an equal representation of all of them. Therefore, since we would have 12 children, we created a list in which we added each skin colour three times and then randomly sorted the list.

The last invalid cue, which is the clothing, had a similar implementation as all the other invalid cues. For the clothing, we had four lists in total (one for branded shirts, one for unbranded shirts, one for branded shorts and one for unbranded shorts). Then each list was randomly sorted. Finally, we had another list where we alternated if the clothes would be branded or not, and randomly sorted that as well. That would indicate if we would draw a branded shirt and a branded short from the respective lists or unbranded ones. The lists were looped so that we would have enough elements in each list for all 12 children.

Moving on to the Animation system, to achieve systematic variation with our animations we had to think of a way to combine all the cues in such a way that all users end up seeing the same amount of cues every time. As mentioned earlier in Section 3.1.2, we had to generate 12 different children (invalid cue variations) and five different animations for each child (valid cue variations). For the valid cue variations, we motion captured 21 animations from two actors (42 animations in total). The process will be explained in Section 4.2. We split the animations into lists so that each different type of animation is in its own list and then randomly sorted each list. Through this, we ended up with six lists of randomly sorted animations. Then we were able to select the first animation from each list, and we would end up with a set of five animations. To get another set, we would get the second animation from each list, then the third and so on. Since not all lists had the same amount of animations, we would restart the selection of animations from a list that ran out from the start.

Lastly, the MoCap variation algorithm applies variations to the positions of the arms and legs on runtime through the OnAnimatorIK(int) Unity function. This is done to achieve some variation in what the users see each time in real-time. The way the variations are acquired in our system is by introducing Perlin noise (an ordered sequence of pseudo-random numbers; Randomness, 2013) to the position of the animations that scales within values of -0.1 to 0.1. We chose these values since anything higher would yield abnormal results and anything lower would introduce seamless alterations. The random Perlin noise would then be added to the limbs of the 3D model.

After the user assesses the digital children, the cues of the different variations along with the corresponding grades from the assessor are stored to be used by the next part of the system. The grades are any value from 1 to 10. The cues needed to be formatted into binary integer values. To do so, a simple coding of the data needed to take place. For the invalid cues, the following applies:

- For the gender, we coded female as 0 and male as 1.
- For the cultural background, we used 0 for foreign and 1 for local cultural names.
- For the skin colour, we used 0 for pale-skinned and 1 for any other of the variations.
- For inclusion criteria, we used 0 for non-inclusion and 1 for inclusion.
- For clothing, we used 0 for non-branded and 1 for branded clothing.

As for the valid cues, we simply counted the number of mistakes that each animation incorporated and represented each animation using the number of its mistakes. As mentioned earlier, more on the animations will be explained later. The grading of the children was stored as one file and the encoded cues were stored as a different file.

4.1.2 Analysis System

The analysis system is responsible for generating the Lens Model of users based on the cues they were tested on and the grades they gave to the digital children. For acquiring the Lens Model of a user, we developed a NodeJS server that is responsible for calling an R script that executes a Lens Model function from the multicon[10] library. This function requires three inputs:

1. The true assessment value, which is the correct and accurate grade of a child

[10] https://cran.r-project.org/web/packages/multicon/multicon.pdf

2. The perceived assessment value, which is what the users graded with each child
3. The cues that were used to get each grade with the binary integer coding

The output of this function is a CSV file with the lens model's coefficients which give the polarity of the bias of a user with regards to each cue. What polarity means, is that if the coefficient is negative, then the user favours the value coded as 0 of the specific cue (from Section 4.1.1). If the coefficient is positive, then the user favours the value coded as 1 of the specific cue. For example, for the cue Skin Colour, a user might get a coefficient value of -0.5, meaning that they favour pale-skinned coloured children since we coded it as 0. The closer a coefficient is to 0, the more balanced the assessor's grading is. These coefficients are then stored, so we can create a full user profile and use them to present the results to our users in the feedback phase.

To acquire the true assessment value of the exercises, we used the grading scheme of the TGMD-3 framework. Two of the researchers of this study assessed the simulations that were applied on a mannequin (Figure 4.1) with no characteristics whatsoever, to not induce bias in their judgment. In the end, the judgment of one rater was compared with the other. In some cases, the grades were different by a large margin (over 0.5 of the grade) or the animations were deemed unclear. Therefore, those simulations were dropped. In total four simulations were dropped out of the study, leaving us with a total of 42 simulations.

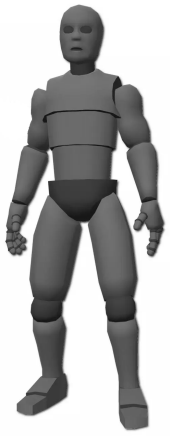


Figure 4.1. The mannequin used for acquiring the true grades of the exercises

4.1.3 Feedback System

For this module of the system, we wanted to find a way of transforming the lens model into easily readable feedback to supply to our users. For that reason, we set some upper and lower bounds for the coefficients that the lens model calculates for the cue utilization of each participant. We wanted to separate the feedback into four levels of importance. To check for the value of the coefficients, we always used the absolute value since the sign of the coefficient did not matter for generating the importance level.

The feedback for the invalid cues was separated into:

1. Average favour of INVALID_CUE
2. Slightly high favour of INVALID_CUE
3. Higher preference favour of INVALID_CUE

4. Extremely high favour of INVALID_CUE

The feedback for the valid cues was separated into:

1. Generally, you grade EXERCISE_TYPE exercises correctly
2. Sometimes you grade EXERCISE_TYPE exercises incorrectly
3. Most of the time you grade EXERCISE_TYPE exercises incorrectly
4. Generally, you grade EXERCISE_TYPE exercises incorrectly

When running tests, we observed that the coefficients' values are mainly between zero and one. The aforementioned boundaries that were set, are based on what we believe a fair distinction between the different levels of feedback is based on possible values of the coefficients. The bounds that we set for the coefficients (x) for the invalid and valid cues are (in corresponding order with the above):

1. x >= 0 AND x <= 0.2
2. x > 0.2 AND x <= 0.5
3. x > 0.5 AND x <= 0.8
4. x > 0.8

For example, if the coefficient of a participant for the invalid cue *gender* is 0.8, then the text on the *gender* 3D bar chart would be *Extremely higher favour of male children*, whereas if it was -0.8, then the text would read *Extremely higher favour of female children.*

Figure 4.2. Feedback and training phase with the invalid cues

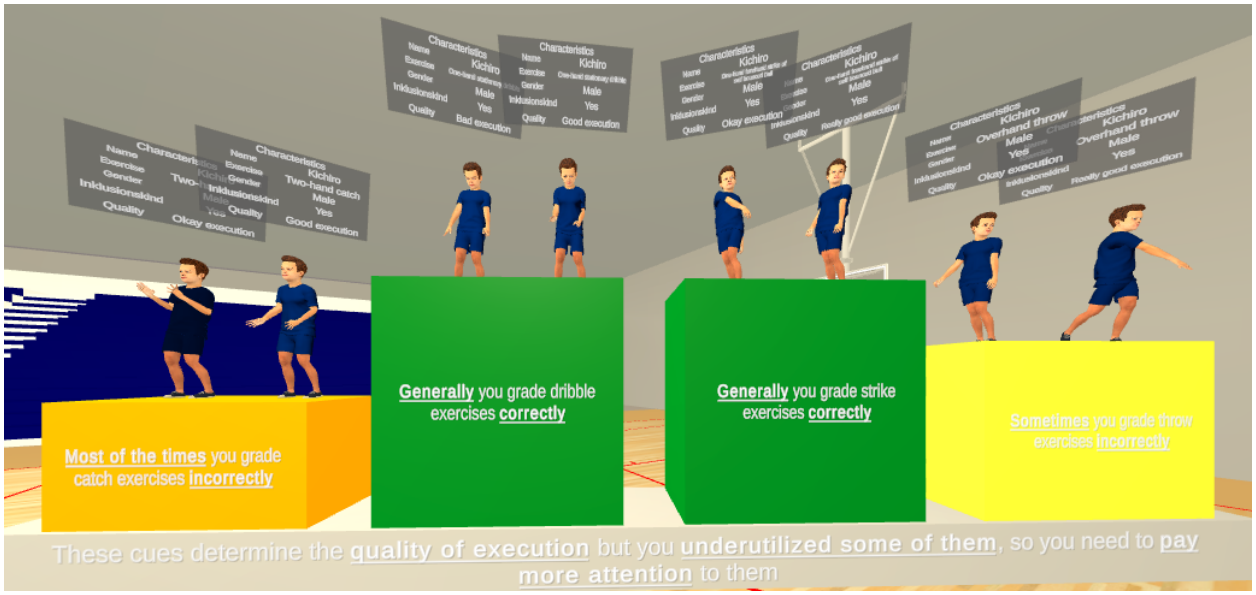


Figure 4.3. Feedback and training phase with the valid cues

For showing the feedback to the users, we wanted a meaningful way of demonstrating the four levels of importance and for the users to be able to compare each cue with the others. For that reason, we decided to use 3D colour-coded bar charts. For the feedback of both invalid and valid cues levels, we used the same colours:

1. Green
2. Yellow
3. Orange
4. Red

The text of each of the cue utilization is imprinted on each of the bars in the bar chart so that users can correlate their feedback to the importance and its representative colour directly. As mentioned the feedback is separated into invalid (Figure 4.2) and valid cues (Figure 4.3). For the invalid cues, we offer feedback for all five of them, and for the valid cues, we offer feedback for four of them since two of the exercises that determine the valid cues are the same type of exercise (Overhand throw and Underhand throw).

4.1.4 Training System

The main goal of the training system is to give some insight to the users as to what they can do to improve their assessment. Furthermore, we want to show to the users the cue they over- or under-utilized to help them understand the output of the lens model. For that reason, during the feedback phase for the invalid cues, we collect one of the digital children of each cue that characterizes their preference and show it to them over the bar chart of the specific cue. For example, if they have high favour of female students, then we would get one of the female student models they graded. Each selected digital child is also performing one of the assigned exercises that it was performing during the evaluation phase. On the other hand, for the valid cues, we retrieve two execution qualities of the exercises that characterize each bar chart of the valid cues feedback. We then show them both side-to-side on one pre-selected random digital

child so that users can compare how the two execution qualities differ, with an informative text of which execution is the better one in the information panel over the 3D model.

Furthermore, for additional feedback, we sum up the feedback importance from the feedback system of a user and display a corresponding message for how to improve their evaluation as the base of the 3D bar charts. For example, if a user, for the five invalid cues, has only green bar charts, then the sum of their feedback importance is zero. For each invalid cue feedback that is not green, we increase the sum by one, therefore having a maximum of five for the invalid cues and a maximum of four for the valid cues. The advice we offer for the invalid cues is as follows:

- If (sum == 0): These cues should be ignored and you did not overutilize any of them.
- If (sum == 1 *OR* sum == 2): These cues should be ignored but you overutilized some of them, so you need to pay less attention to them
- If (sum == 3 *OR* sum == 4): These cues should be ignored but you overutilized most of them, so you need to pay less attention to them
- If (sum == 5): These cues should be ignored but you overutilized all of them, so you need to pay less attention to them

We used the same advice in some of the texts since we wanted to increase the importance of the advice with the same scale as the feedback bar charts.

As for the valid cues, the advice we offer is as follows:

- If (sum == 0): These cues determine the quality of execution and you did not underutilize any of them.
- If (sum == 1 *OR* sum == 2): These cues determine the quality of execution but you underutilize some of them, so you need to pay more attention to them
- If (sum == 3): These cues determine the quality of execution but you underutilize most of them, so you need to pay more attention to them
- If (sum == 4): These cues determine the quality of execution but you underutilize all of them, so you need to pay more attention to them

4.2 Motion Capture

For performing the motion capture, two actors were recruited. As mentioned earlier, we used the OptiTrack Prime 13 hardware. The actors had to perform each of the aforementioned exercises from the TGMD-3 framework but in multiple iterations. The initial iteration was performed with no mistakes introduced into the exercise and each subsequent iteration had one mistake introduced into the performance while keeping the previous execution quality. For example, for the exercise Overhand Throw this is the acting order:

1. Performance with no mistakes
2. Performance with the first mistake (Windup is initiated with a downward movement of hand and arm) introduced
3. Performance with the second mistake (Rotates hip and shoulder to a point where the non-throwing side faces the wall) introduced while also having the previous mistake
4. Performance with the third mistake (Steps with the foot opposite the throwing hand toward the wall) introduced while also having the previous two mistakes

5. Performance with the fourth mistake (Throwing hand follows through after the ball release, across the body toward the hip of the non-throwing side) introduced while also having the previous three mistakes

After acquiring the animations, we then had to perform cleanup by relabelling the body parts that were not labelled correctly as well as removing or fixing frames in the animations where the capture quality was not good enough. In some cases, animations had to be discarded, which led to having multiple recording sessions.

The process of motion capture is demonstrated in Figure 4.4 where the reader can see the actor performing the overhand throw exercise in the first segment, followed by the outcome in the MoCap software that we used to clean up the MoCap, followed by the digital child when the animation is applied to it, and lastly when a digital child is in the VR environment while performing the animation.

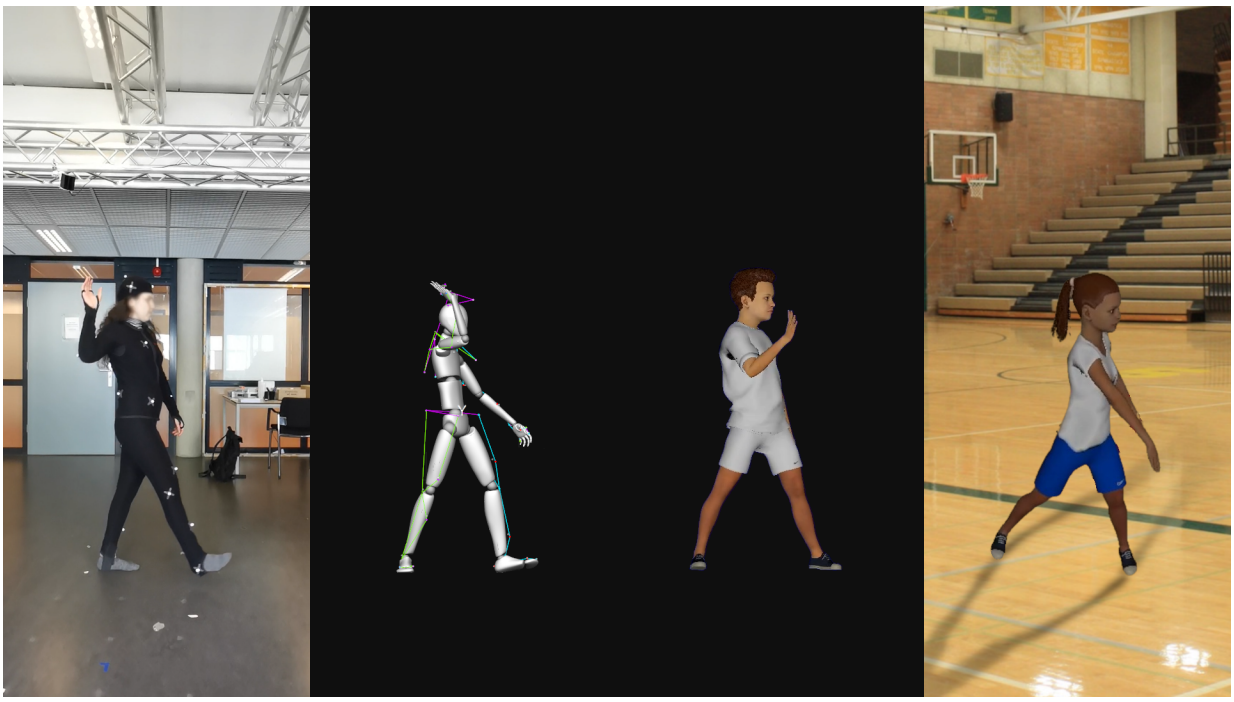


Figure 4.4. The evolution of our system from MoCap to the VR system

4.3 Children Models

Earlier in this chapter, we discussed how we varied each invalid cue on the digital children. This part of the chapter will discuss how we acquired the model of the children as well as the clothing. As discussed, we used the o3n UMA Races Stunner Jane & John Standard Render Pipeline. This asset allowed us to have a male and female model on which we can vary characteristics, such as age, height, etc. For our case, we changed the age slider to the minimum setting, so it produced young children, but we also lowered the height setting because we felt like the children were a bit too tall in the simulations. Furthermore, this asset gave us the functionality of applying clothing to different body parts of the models by applying tags to the clothing models (Pants for anything that is worn on the legs, Chest for anything that is worn as a

shirt, Shoes for shoes, etc). Through this, we could change the shoes and hair of the models to some pre-existing models that the asset already provided us with. We used a ponytail model for the female hair and a short-cut hair model for the male hair. For the shoes, we used the sneakers model that was available.

For the clothing 3D models that we would wear on the digital children, we 3D modelled shirts and shorts, as well as applied textures and colours to them, with the help of a colleague who is experienced in 3D modelling and therefore could achieve a realistic result quickly. As mentioned earlier we created branded and unbranded pairs of clothing, and more specifically four unbranded and four branded pairs. The models are the same for all eight shirts and eight shorts, and the only thing that varies is the colour (white, black, blue, red) and the branding. The clothing was created and rigged in Blender.

Moreover, for the skin colour, we wanted to create some variety. For that reason, we based our selections on the Fitzpatrick's skin colour scale (Fitzpatrick, 1975) for skin tones of type III, IV, V and VI (Figure 4.5).

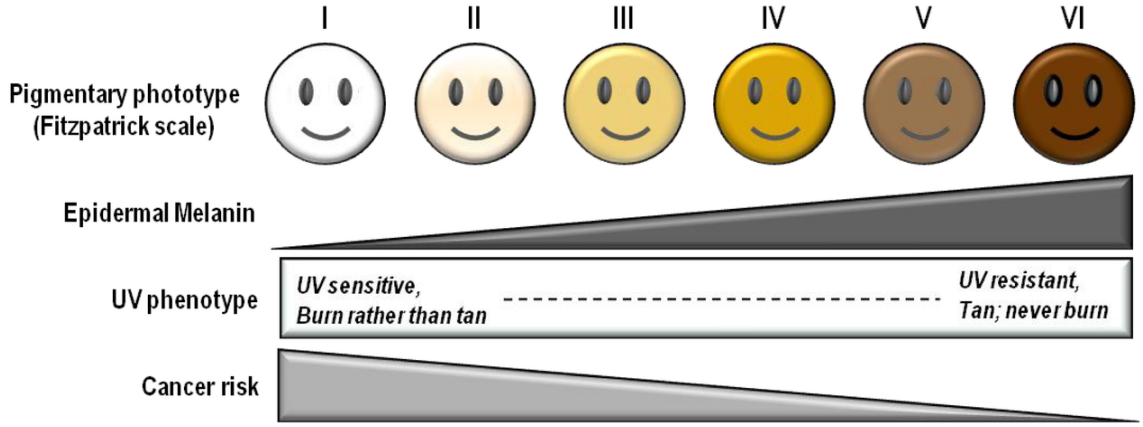


Figure 4.5. Image from Wikipedia [11]which shows the skin colour types based on Fitzpatrick (1975)

Lastly, in our simulation, we wanted to have the ball that the digital children are throwing or grabbing during the exercises. Motion capture was performed on a physical ball while the exercises were performed by our actors to acquire the ball's movements. Due to technical difficulties though, we needed to simulate the physics and movements of rigid bodies along with other animated objects. Because of the time constraint of the project, we could not do so and we decided to remove the ball from the simulations. Therefore, we acknowledge that the ball is a missing element of the simulations.

4.4 Environment & UI

Related to the environment of the simulation, in this subchapter we will discuss how users can interact with it and the UI.

First and foremost, moving around in the environment can be done using two methods. The first one is by physically moving around in the environment. Since virtual reality translates the physical position of the user from the starting point of the simulation to in-game movements,

[11] https://en.wikipedia.org/wiki/Fitzpatrick_scale

users can physically walk around the environment. As an alternative method of movement, we added teleportation. Users can hold down the index trigger of the left controller and point anywhere on the ground where a circle appears on the digital ground. By letting go of the trigger, the users teleport to the highlighted position. Furthermore, for rotating in any direction, by moving the right-hand joystick to the left or the right, the view in the simulation rotates to the respective direction by 30 degrees. This allows for more freedom while moving using controllers, without the fear of running out of space in the physical world. Moreover, we added a bounding box around the basketball court of our simulation so that users cannot freely walk out of the simulation's space.

As for the UI, users can point the right controller on any of the UI elements, where a purple line appears from the controller to the UI element, signalling an interaction possibility. For clicking a UI element, they can press the right index trigger while pointing at a UI element. This allows them to use the left-hand media key panel as well as inputting a grade using the input field. When users click on the input field, a digital VR numeric keyboard appears which they can use the right-hand controller to interact with.

4.5 Computer Control Version

For developing the computer control version of our system, we used Angular 12. We created a website with two child routes, one for the evaluation system and one for the feedback system. The evaluation page, as mentioned earlier in Section 3.6, is playing videos of a VR simulation that was recorded from four different angles. This is to give to the users a somewhat 360 degrees perspective view of the model. We added media keys for skipping to the next or previous exercise, as well as pausing and playing the animations. Moreover, users can input their grades in an input field and submit their assessments so they can progress to the next child. We tried to give as many similar controls as possible with the VR version. The processing of the cues and grades is the same process as the VR system.

As for the feedback phase, the data processing that we used is the same as described in Section 4.1.3. We followed the same structure and principles which yielded the result as shown in Figure 3.12. Our goal was to keep the feedback phase on the computer as similar as possible to the VR one. Therefore, we used the same colour scheme and the same amount of information and information placement. The only difference is the lack of animated models that display the valid and invalid cues as described in Section 4.1.4.

4.6 Server

For the server-side of things, we used NodeJS.

We implemented 4 endpoints:

1. For generating and saving the cue utilization of the users in VR
2. For generating and saving the cue utilization of the users on the computer
3. For retrieving the cue utilization of users in VR
4. For retrieving the cue utilization of users on the computer

The first two endpoints, perform a call to an R script, which is executing the Lens Model function from the multicon library as mentioned above. As soon as a call is executed to one of the endpoints that are responsible for generating the cue utilization, a folder is created in the server

with all the details that are needed for modelling the user's bias. These details are the cues that they were exposed to, their assessment and the true assessment of the simulation they assessed stored as CSV files. When their utilization is generated, a CSV file is generated and stored in the same folder. The final result is a folder with each user's profile and their qualities as an assessor.

Chapter 5 - **Evaluation**

This chapter will focus on the user study we conducted to perform a summative evaluation of our system. This section contains two subchapters that focus on the design of the study and the study itself. One remark is that we aimed to conduct a second user study for the computer control version. However, the study was conducted in August, and most of the experts and students of our target group were unavailable. Therefore, we only conducted the within-subjects study of our original planning.

5.1 Study

For evaluating the system, in its current form, only qualitative data could be collected. The study consists of a within-subjects experiment where the users are exposed to both feedback phases in VR and on the computer in an alternating fashion. This means that when a participant is exposed to the VR feedback first and then to the computer feedback, the next participant will be exposed to the computer feedback first and then to the VR too.

The collection of data was conducted through two media. The first was Likert-scaled open-ended questions (questions with a five-point scale). The second was a semi-structured interview with follow-up questions. The structure of both can be found in Appendix B and C respectively.

5.1.1 Procedure & Materials

The study took place on three consecutive days, in two rooms at the University of Muenster. The duration was estimated at 1 hour and 30 minutes, with a participant every 45 minutes in each room respectively. The first participant of every day was scheduled for 9 a.m. and the last for 3.45 p.m. The study procedure with each participant was the following:

1. The participant was welcomed to the study and requested to wash their hands with disinfectant due to COVID-19 regulations and wear a mask if they did not already (both disinfectant and mask were provided by us)
2. The participant was given a COVID-19 health declaration form which states that they do not have any symptoms
3. The participant was given the informed consent form (Appendix D)
4. After reading the information brochure the researcher answered any questions that the participant may have had before they signed the consent form
5. After the participant signed the consent form, they were briefly introduced to the experiment again and given the chance to ask questions
6. The researcher requested from the participant to stand in the centre of the play area and help the participant put on the controllers
7. The researcher briefly explained to the participant the controls so that the participant would understand what they will be doing and how
8. The researcher helped the participant put the HMD on

9. The researcher explained the controls again while the participant was performing the actions for the participant to train with the controls
10. The researcher allowed the participant to train using the controls for a brief period of time until the participant was confident using them
11. When the participant was comfortable with the controls, the researcher informed the participant that they would start recording them and the participant could start the simulation
12. When the participant was done with the simulation, they were exposed either to the computer feedback or the VR feedback first and then to the other device (according to their entry in the within-subjects schedule)
13. The recording was stopped and the participant was requested to fill in the questionnaire (Appendix B)
14. The participant was interviewed (Appendix C)
15. The participant was debriefed and thanked for their time

Figure 5.1. One of the participants while performing the simulation

The equipment needed for conducting the study was borrowed from the HMI Lab. The necessary equipment was:

- 2 x Oculus Quest 2 and power banks for prolonged battery life
- 2 x Camera for recording the participants while performing the experiment
- 1 x Laptop for the participants to fill in the questionnaire, reviewing their feedback on the computer and writing their interview answers on
- Cleaning supplies for following the COVID-19 regulation

The setup of the study can be seen in Figure 5.2. Both participants were in two separate rooms, with a camera filming each participant. The space that the participants had free to physically move around was approximately two meters by two meters. The researcher was

moving between rooms for the participants to review their feedback on the computer, fill in the questionnaire and participate in the interview.

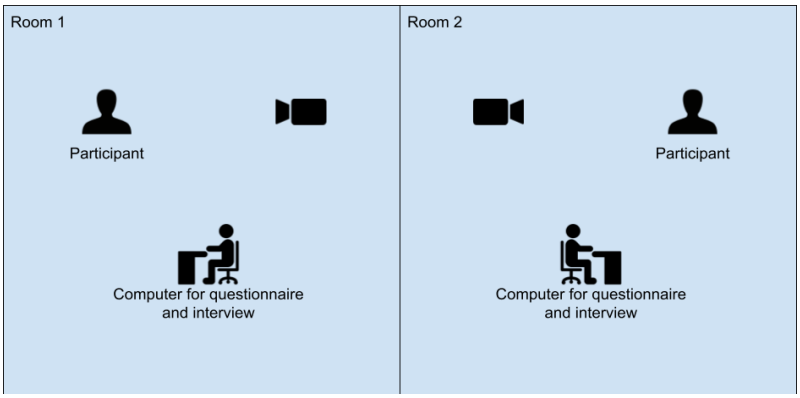


Figure 5.2. Study setup

5.1.2 COVID-19 Regulations

To ensure the safety of both participant and researcher, specific COVID-19 regulations were in place. The regulations were approved by the Ethics Committee of the Faculty of Electrical Engineering, Mathematics and Computer Science of the University of Twente.

The participants were requested to put on disinfectant as soon as they entered the study room, and fill in a health declaration form that was supplied by the University of Twente. If the participants were not already wearing a mask, the researcher had available masks to provide the participants with. Moreover, they were briefed about the COVID-19 regulations that would be in effect throughout the study.

The researcher was also responsible for disinfecting all the devices that would be touched by all participants, such as the HMD, controllers, pens and the computer device after each use. Furthermore, a protective one-use mask was applied on the HMD where the headset is attached to the user's face. That protective mask was replaced after each participant was done with the experiment. In addition, windows were open to allow for clean air to circulate through the rooms. Unfortunately, the 1.5 meters distance rule was not always feasible to follow, since participants required help to put the headset on as well as taking it off and in case of motion sickness, they had to be assisted to a chair for a cooldown. Throughout the study, we tried to minimize the occurrences of cases where the distance rule would be violated.

5.1.3 Participants

Since our project's goal is to educate people regarding their bias on primary school PE exercises, our target group is mainly PE teachers and PE teachers-in-training. Since the external supervisor of this thesis project is located at the University of Muenster, where they have facilities for training PE teachers, our participants were drawn from that university. Furthermore, since we wanted to see the overall perception of people for their bias, and due to summer holidays, we also recruited psychology students and teachers-in-training of other courses as well, such as biology and chemistry.

In addition, we had two additional recruitment criteria for the study. Participants needed to be competent adults, meaning adults who would be able to complete the study due to its complexity since we are working with a niche technology. The second was that participants needed to be English readers and speakers since the entirety of the study is in English and we wanted to have direct answers in English without the need of translating items and losing the quality of our data in the translation.

The recruitment process was handled by the external supervisor, where they asked for willing participants. In total, we recruited 30 participants, from which four cancelled, one experienced motion sickness and technical difficulties and stopped the study, and one was removed due to technical difficulties, giving us a final total of 24 participants (female = 13), with 19 being in their Bachelor's, two in their Master's and three currently in other positions in the education domain (Figure 5.3). Furthermore, nine of our participants had prior experience in VR with one of them using a VR headset at least once per month.

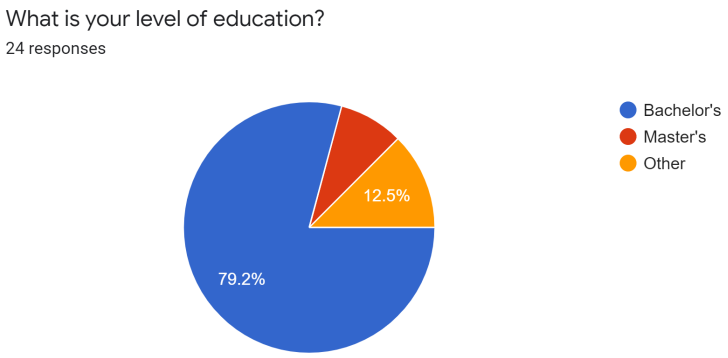


Figure 5.3. Participants' level of education

5.1.4 Measurements

Since one of the main focuses of our study is whether our system works and is appealing to the users, we decided to conduct a usability questionnaire (Appendix B) to test the user experience of our participants. The questionnaire we adopted is the User Experience Questionnaire[12] (UEQ) which is widely used in many cases which need to evaluate the user experience of a product (Schrepp et al., 2014). The questionnaire aims to evaluate a product with the use of 26 adjectives. We removed some of the adjectives that we did not see fitting in our experience and we used a 1 (completely disagree) to 5 (completely agree) scale to allow the participants to rate their experience. More specifically, the UEQ tests a product on six scales:

1. Attractiveness: whether the participants liked the system or not
2. Perspicuity: whether the system was easy to get used to and learn how to use
3. Efficiency: whether the participants could perform their assigned task without too much effort
4. Dependability: whether the system feels secure and predictable
5. Stimulation: whether the system is exciting and motivating to use

[12] https://www.ueq-online.org/

6. Novelty: whether the system catches the interest of the participants and the design is creative.

In our case, due to our selection of adjectives, we did not test dependability. Furthermore, we wanted to better understand how participants perceived the feedback phases of both VR and the computer, as well as what they thought of the additional feedback elements that we presented in VR. To achieve that, we added Likert-scaled and open-ended questions targeted at the specific theme.

In addition, we wanted to test what the participants thought of the experience overall and their perception of how we simulated the animations and the 3D models. Furthermore, we wanted to see whether a conversation could be kickstarted by the participants to talk about their feedback. To do so, we conducted semi-structured interviews where we collected user quotes and asked clarification questions about the participants' experiences.

Lastly, quotes and behaviour patterns that the participants expressed during their time of using the system were extracted from video recordings to have more insights on their questionnaire answers, as well as kickstart more discussion during their interview.

5.2 Analysis

To acquire insights for the Research Questions we set to answer, we will discuss how we analysed the data we acquired through the questionnaire, interview and video recordings that were discussed in the previous subchapter.

Starting with the questionnaires, and more specifically the UEQ questions, the results yielded inconsistencies in a question related to the Novelty scale. We were able to calculate them through the tools the UEQ provided us with, and that specific question was removed from the data. Furthermore, we flipped the values of the negative-shaped questions and adjusted the range of the answers to -2 (strongly disagree) to 2 (strongly agree). We then calculated Cronbach's Alpha value (how closely related a set of items are as a group; Tavakol & Dennick, 2011) to understand the validity of the themes that we were looking to investigate. You can view the alpha values in Table 5.1. Based on the results, we can accept the answers related to the Attractiveness, Perspicuity and Novelty scales as reliable, but not the Efficiency and Simulation scales (accepted Cronbach's alpha values need to be over 0.7).

Scale	Alpha (a)
Attractiveness	0.84
Perspicuity	0.77
Efficiency	0.63
Stimulation	0.40
Novelty	0.80

Table 5.1. Cronbach's Alpha result for the UEQ results

With regards to the rest of the questions in our questionnaire, no specific analysis processes have to be performed since we will use them as descriptive statistics later in our results (Chapter 6) to answer our research questions.

When it comes to the video recordings, we extracted quotes and behaviours of our participants while using our system. The behaviours we were looking for were any form of movement apart from teleportation around the environment and filling in the grade. As for the quotes, we did not record any questions related to how to use the system.

Lastly, for processing the data we collected from the interviews we used a top-down coding strategy. The first step was to create the themes based on the research questions we set in Section 1.2, and for each theme, we created several categories (See Table 5.2).

	Themes				
	Feedback	Bias	Animation/ Model	VR	General Comments
Categories	General	Accepting	Repetition	Likeability	
	Interface	Denying	Uncanny	Interface/Hardware	
	Understanding		Likeable		
	Usefulness		Realistic		

Table 5.2. Themes and categories of the interviews' analysis

One coder reviewed the interviews before initiating the coding and then placed the codes in the related categories and themes. We then reviewed the findings to draw conclusions and understand how the participants perceived and experienced the assessment tool. With the interviews and the think-aloud protocol we also wanted to see the number of people who initiated a discussion about their feedback and bias to understand how many of our participants were interested in their results.

5.3 Results

The study was aimed towards answering the third and fourth research questions, namely:

- RQ3: How well perceived and appreciated is the developed system?
- RQ4: How much does the VR feedback phase add to the 2D version of the feedback?

In chapter 1.2, we explained in more detail what we are looking to investigate through our research questions. As a refresher, we want to see what is the response of the participants to the VR assessment and the resulting Lens Model. Moreover, we want to see if sensible discussions can be initiated by the participants to talk about their feedback. As a result, we would like to see whether our system is fit for modelling the participants' bias and if the participants learn from it. In addition, we want to find out whether the VR aspects that we

incorporated into our solution help in doing so, and what are the resulting strengths and weaknesses of our solution in that regard.

5.3.1 Questionnaire

5.3.1.1 UEQ

Based on the analysis of the data we extracted for the UEQ questionnaire, we present the medians, distributions and outliers (empty circles and stars with numbers, which indicate extreme values that fall out of the boxplots) of the participants' answers in Figure 5.4. In addition, the means and standard deviations are presented in Table 5.3. It is noted that all means are over 1 (agree) with Perspicuity being the highest, followed by Novelty, then Stimulation, Efficiency and Attractiveness. The scoring of all aspects being over 1 is promising since it indicates that participants rated that the system has a good user experience overall.

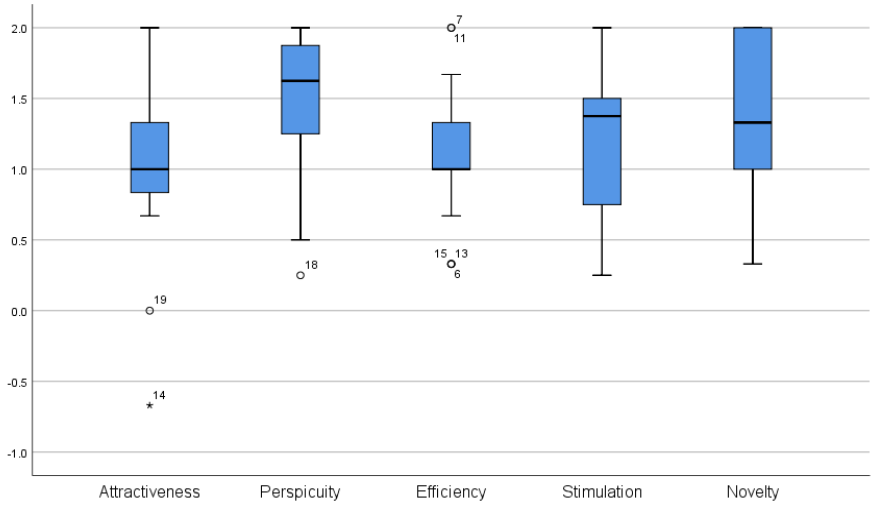


Figure 5.4. Boxplots of UEQ questions

	Mean	Std. Deviation
Attractiveness	1.028	0.573
Perspicuity	1.521	0.477
Efficiency	1.097	0.446
Stimulation	1.188	0.462
Novelty	1.389	0.517

Table 5.3. Mean and Standard Deviation of UEQ answers

5.3.1.2 Experience

As noted earlier, we asked further questions regarding the experience, as well as clarifications about the feedback phase on both VR and computer. In Figure 5.5 we present the medians, distributions and outliers of the answers to those questions and in Table 5.4 their means and standard deviations. We observe that the mean of participants needing technical support for using the system is close to 0, and participants did not experience inconsistencies in the system. Furthermore, they rated the task as repetitive, with sufficient animation controls and a communicated goal. The answers to the question related to the need for technical support was surprising since VR overall is not a well-known and well-distributed system. On the other hand, the response to the repetition question was expected since, although there is variety in the models, the task itself is repetitive.

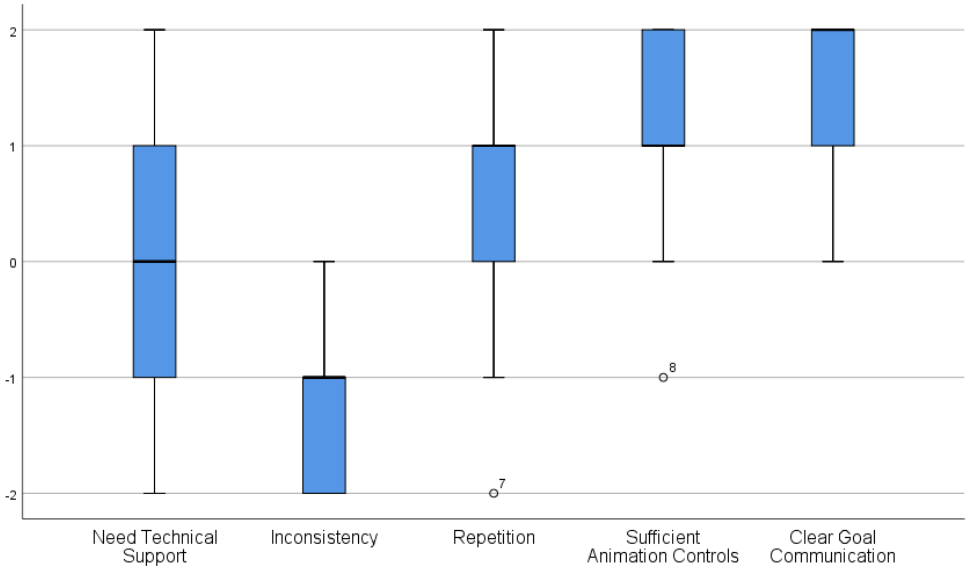


Figure 5.5. Boxplots of other usability questions

	Mean	Std. Deviation
Need Technical Support	-0.042	1.233
Inconsistency	-1.292	0.624
Repetition	0.667	1.090
Sufficient Animation Controls	1.167	0.816
Clear Goal Communication	1.500	0.722

Table 5.4. Mean and Standard Deviation of other usability questions

5.3.1.2 Feedback

The feedback phase was the next aspect of our system that we evaluated through our questionnaire with both Likert-scaled questions and open-ended questions. When the participants were asked how valuable they found the feedback phase on VR and computer, they rated the VR feedback phase more valuable overall (Figure 5.6). We expected to see some equality in the results regarding this question since computer devices are more common and participants are more used to them.

In addition, for the questions targeted at the feedback phase on VR (Figure 5.7), participants indicated that it was easy to understand and the colours that were assigned to the 3D bar charts helped understand the feedback. Moreover, the 3D models on top of the 3D bar charts helped understand the feedback, as well as the animations with the different execution qualities they were performing for the valid cues. When asked whether the VR feedback phase convinced them more than the computer feedback phase, the mean response of participants is 1 (agree) with an extreme outlier with a value of -2. These responses hint us to believe that the structure of our feedback phase was a good approach to the problem.

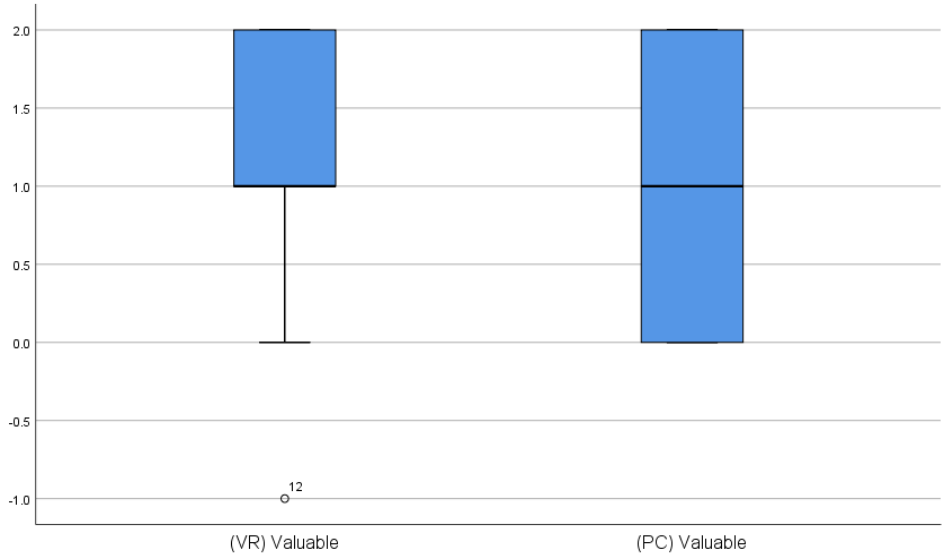


Figure 5.6. Boxplots of answers to whether the participants found the VR feedback and the PC feedback is valuable

	Mean	Std. Deviation
(VR) Valuable	1.208	0.779
(PC) Valuable	1.000	0.834

Table 5.5. Mean and Standard Deviation of answers to whether the participants found the VR feedback and the PC feedback valuable

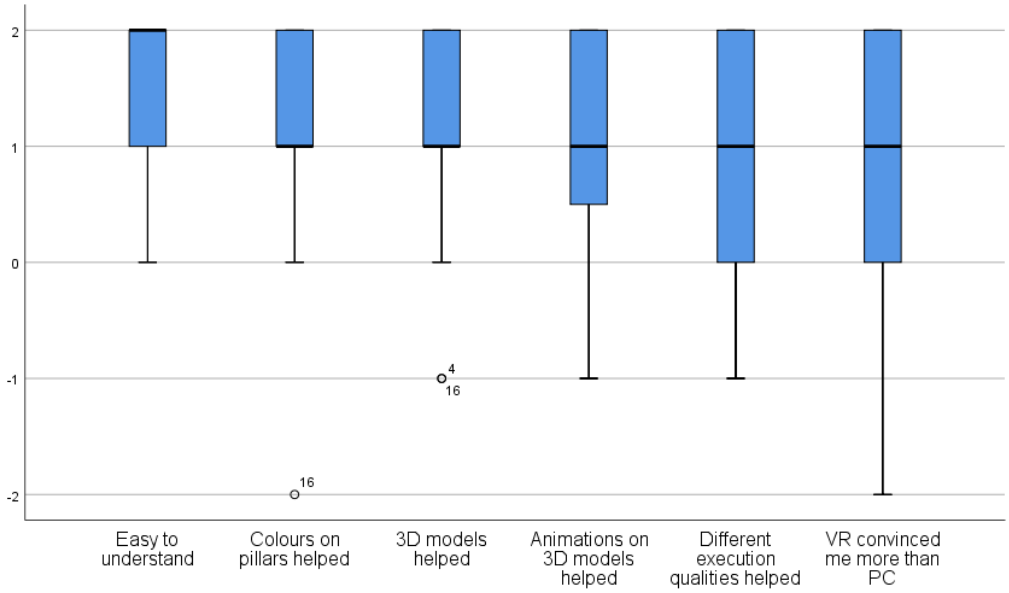


Figure 5.7. Boxplots to answers to VR feedback phase targeted questions

	Mean	Std. Deviation
Easy to understand	1.500	0.590
Colours on pillars helped	1.125	0.992
3D models helped	1.208	0.884
Animations on 3D models helped	0.875	1.116
Different execution qualities helped	1.000	1.063
VR convinced me more than PC	0.708	1.334

Table 5.6. Mean and Standard Deviation of answers of VR feedback phase targeted questions

To supplement the Likert-scaled questions, one open-ended question was asked to see whether the participants thought the purpose of the feedback phase on the two devices was different and if yes, how so. Eight participants either responded negatively or said that they “don’t know”. Ten participants stated that the VR feedback phase had a purpose to show the movements again and/or make it easier to understand because they could see the motions again to get more information. More specifically a participant stated “In the VR at least I could actually see what was the good and what was the bad way of doing the exercise. Also maybe the purpose was to show you your results in a big way for realisation purposes.” and another “in the vr I have been able to see both ways of execution of the motions. This was helpful to

understand whether my assessment was rather good or bad.” One of the ten participants, also stated that “you get the feedback in a new form (3D)”. In addition, a participant stated that “it's more impressive in VR. I felt more distanced (passive) to watch the information on the computer. I think I will remember the results better from the VR simulation.” while another replied with “The feedback on the computer is just a real quick glance at the score, the feedback in VR was more fleshed out and made me really think about what I've done because of the animated children and the way it was presented.”

On the other hand, a participant noted that the purpose of the feedback was different because “[...] there are different sentences presented”. Another response was that the computer helped them more since it provided statistics from a visual point. This response hints at the use of bar charts in the feedback phase on the computer to demonstrate their bias. Finally, one participant noted that the purpose was not different, but “[...] maybe in VR you could see the differences of evaluation more clearly”.

5.3.1.2 Overall

The final question was targeted to understand the preference of users towards the feedback device so that we can get their overall preference of the feedback device (Figure 5.8). The majority of participants (20 in total) selected VR as their preferred feedback device. This is a promising response from our participants towards uncovering whether VR is a viable solution for providing training to people.

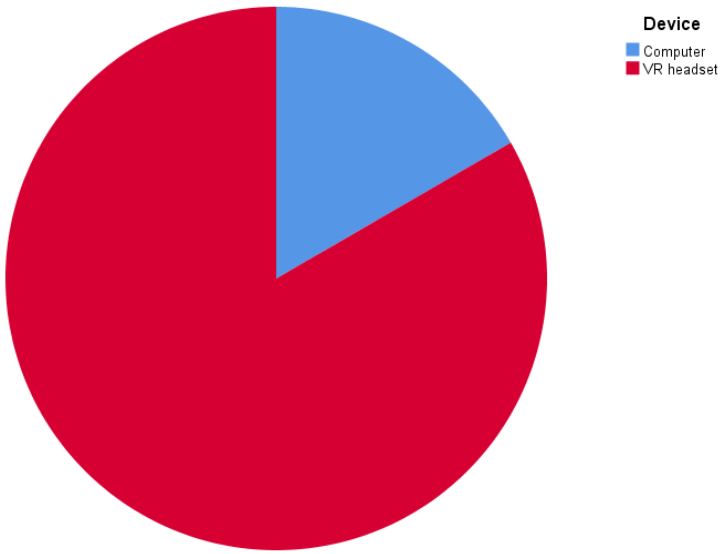


Figure 5.8. Participants’ preference of feedback device

We asked a question to supplement the choice of their preferred feedback device. One of the four participants who selected the computer responded that it was because they are used to using a computer to look at the information. An important note is that another participant said that it is sufficient to look at the feedback on the computer and that “wearing VR and being in there for me is too immersive and mentally exhausting”.

For the participants who prefer VR, the responses talked about the experience being more visualized and easier to understand. In addition, it demonstrates your mistakes and it is

easier to memorize them. Furthermore, a participant noted that it “Made it feel more real or personal and spoke to me on a more emotional level than just basic numbers or charts.”

5.3.2 Video Recordings

The video recordings, although not fruitful in terms of quantity of results, provided some interesting insights on the behaviour of our participants as well as some quotes while they were thinking aloud while performing the study.

In terms of movement during the evaluation, two participants were crouching frequently and when they were asked why they were doing, they stated that they wanted to observe the models from a closer perspective. Another participant started imitating the exercises that the 3D children were performing to see how they would perform them. On a different note, a participant laughed while assessing a child and they said “I don’t know this exercise, it looks funny”. Finally, a few participants asked whether they could fix the grades of other children they graded earlier in the simulation when they progressed through the assessment.

5.3.3 Interviews

In this subchapter, we will mention the results of every theme (Table 5.2) that was extracted from the interviews.

The first theme is Feedback. In total 11 participants initiated a conversation or discussed the feedback they received from the system. Generally, participants said that it was interesting to see how they performed and that they could discover their prejudice, but also because they can learn from it. Regarding the interface, they liked the colours on the 3D bar charts, but also the feedback phase, in general, felt closer to them, and more specifically “it’s as if you are in the situation”, as a participant stated, than the computer feedback. On the other hand, a participant mentioned that the yellow colour of the 3D bar charts made it hard to read the white text that was imprinted on them. Furthermore, a participant specifically said that “It was easier for me to understand my feedback”. In addition, a participant asked the researcher for further advice on how they can fix their assessment based on their feedback. Lastly, participants considered seeing the children during the feedback to be helpful. An example is a participant who was not aware of the branded clothing cue on the computer, but in VR it was much more clear since they could see the cue on the digital children.

With regards to the Bias theme, some participants expressed acceptance towards their calculated bias mentioning that they were shocked by their results. More specifically, one participant realised that they favour children with pale skin tones and commented “oops okay that’s bad”, while also stating that “it's a good study because I want to be a teacher and this made me think about my grading and I was shocked when I found out that I grade white children extremely better”. Another participant throughout the study felt like “the difference of the black and white children was so much that I thought that the black children were performing better, I think the black person performed generally well but the white person performed in general worse”. On the other hand, some participants expressed denial towards their bias. One stated that “The brand of the t-shirts because i did not think about them because the experience says that i looked at it but i tried not to look at it”. Another mentioned that they did not notice the skin colour in the experience, although their lens model calculated that they favour pale skin colours.

We need to clarify that although we are confident we have modelled the participant's bias correctly, we cannot say so with absolute certainty. In particular, in Figure 5.9 we present the box plots of the participant who denied their pale skin colour bias. Based on their grading, they graded children with pale skin colours (mean = 6.7, stdev = 2.507) higher than darker-skinned (mean = 5.3, stdev = 2.087) children, which is an indication that the real bias might have been calculated. In addition, some participants discussed that facial expressions and motivation play an important role in what grade they assign to children. A participant mentioned, "I kept reminding myself that it's a computer and it's not a real person and doesn't really take emotions into account". Furthermore, only a few participants said that they did not pay attention to the characteristics panel over the children's heads. Lastly, a participant was grading due to their personal experience that girls perform better than boys since they are more motivated.

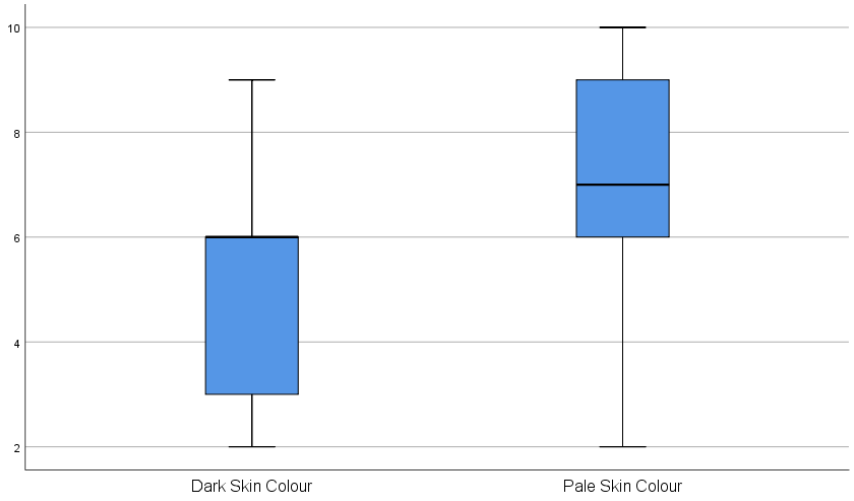


Figure 5.9. Boxplots of a participant's skin colour cue grading

The next theme was Animation/Model. In total, when participants were asked how many actors they believe we performed MoCap on, on average, they answered 8.65 actors with the highest answer being 60 actors and the lowest being 1 actor. In Figure 5.10 (mean = 8.652, stdev = 12.115) we present how many actors the participants believed we recruited. In addition, 22 participants stated that they noticed repetition in the animations and one was inconclusive. In general, regarding repetition in the simulation, many participants stated that some animations were similar while a few said that some children looked the same but the names were different. Throughout the interview, no participant mentioned anything about the system inducing uncanny feelings. In terms of likeability, some participants expressed that the animations were nice to see, while some preferred digital children instead of actual children "So you can actually concentrate more on the cues and not be distracted by something happening". Almost all the participants expressed that they did not like the fact that the ball was missing from the animations. When it comes to how realistic the simulation felt, many participants expressed that the movement was realistic and impressive and it allowed for free movement with no restrictions when compared to real movement. Furthermore, multiple participants stated that the lack of facial animations and emotions made it sometimes look unrealistic since emotions and motivation play an important role in assessment.

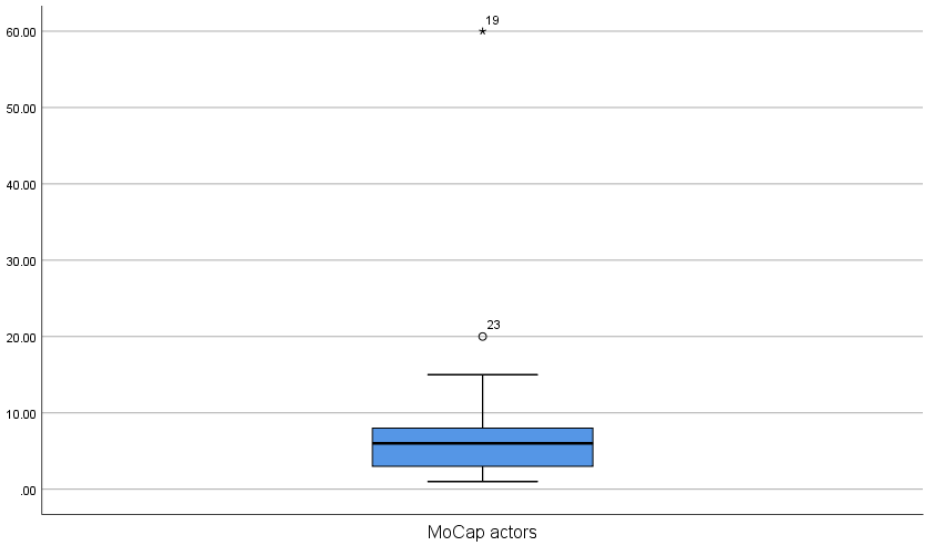


Figure 5.10. Boxplot of how many actors the participants believed we filmed

The next theme is VR, where most of our participants were surprised and enjoyed the simulation, despite the lack of experience in the majority of our participants. Another note is that a few participants disliked the fogginess of the lenses. This was out of our control and caused by wearing masks and the participants' breath ending on the lenses. A participant mentioned that "If you see it on the computer it's not that emotional, it's not that real", but another said that it was hard to stand in one place and their neck and back felt stiff, and their hands started hurting after a few minutes. When asked about the interface and tools that we provided them with, many participants said that it was much easier to rotate the children with the media keys. Others said that they liked the fact they could go around the child and see it from different perspectives. One participant specifically stated that "In a video you only have one perspective and here you have a 360 degrees perspective". A surprising statement was from a participant who mentioned that they cannot use controllers at all in other consoles, but our system was really easy to operate with the controllers. On the other hand, a participant stated that "I thought the refresh rate of the frames was too low".

One of the most often observed comments was that participants would like to have a baseline before the simulation. This would help by establishing a baseline for how an exercise should be performed. On another note, a participant stated that they enjoyed grading children in an environment where the movement was fitting. On the other hand, a participant found it boring after the fifth or sixth simulation and stated that "I could imagine that the results of my evaluation for the children begins to be less accurate". Lastly, some participants discussed that they would prefer to have more children in the simulation and see how children interact with one another to make it even more interesting.

Chapter 6 - **Conclusions**

Throughout this report, we discussed the design, implementation and evaluation of our Virtual Reality system that aims to help teachers and teachers-in-training to uncover their biases, provide them with feedback and educate them. Through the background research (Chapter 2), we have acquired the answer for RQ1: *What are the underlying theories and models of our project?* Throughout the rest of the report, we concluded answers regarding the other three RQs we set: RQ2: *What is the reference architecture of such a system?*, RQ3: *How well perceived and appreciated is the developed system?* and RQ4: *How much does the VR feedback phase add to the 2D version of the feedback?* This chapter concludes the answers to our RQs, followed by a discussion related to the results and research, and closes with the limitations and future work surrounding our project.

6.1 Conclusions

RQ1: What are the underlying theories and models of our project?

a. How is Brunswik's Lens Model used?

The Lens Model is an approach to model someone's bias. To do so, three vectors are required to be imported into the Lens Model's multiple regression. The first vector contains the true measures of the environment. The second vector contains the cues that a person (for whom we want to calculate the LM) uses to perceive the environment. The third vector contains the overall cues that are used to describe the environment. The output is a vector with the weight of the coefficients of each cue, signifying the cue utilization of the person we are interested in (Fiedler, 1996; Douglas & Gifford, 2001).

b. How is Brunswik's Lens Model applied to teachers?

In literature, Cooksey & Freebody (1987) used 5 cues to see their effect on how teachers assess the reading skills of kindergarten students, concluding that their judgement affects their behaviour in the class. Another study used a set of cues to model the perceived understanding of students on a specific mathematical concept by their teachers (Oudman et al., 2018). The results indicated that only knowing the name of a student is enough for teachers to rate the understanding of a student on a subject. Lastly, similar to our approach, Utesch (2020) has generated bias models of teachers when assessing videos of primary school children performing GMD related exercises.

c. How is Interaction Technology combined with Sports?

I-TECH has already been implemented in the case of tennis, for example, the Cyclops system, or for bowling, such as the foul-line detector. Related to our case, mobile system tools to help PE teachers perform fair grading to students have been proposed, as well as the use of videos to understand the assessment biases of PE teachers (Utesch, 2020). The same approach for providing feedback to the students related to their performance has been researched by Preuschl et al. (2010). On the other hand, feedback systems for providing the

teacher with how students perform, have not been extensively researched in literature (Cushion & Townsend, 2019; Castelli et al., 2015).

d. How can MoCap data be varied?

A few methods have been proposed in the literature for achieving variation of MoCap data from a minimized dataset of animation through probabilistic datasets of poses (Boukhayma & Boyer, 2017) and environment processing (Pan et al., 2010). In addition, combining animations were discussed to achieve a smooth result (Boukhayma & Boyer, 2018). In our case, a simpler approach took place where we used a Perlin noise function with a scale of -0.1 to 0.1 that provided some additional movement on the limbs (as explained in Section 4.1.1).

RQ2: What is the reference architecture of such a system?

a. How are its submodules defined?

The reference architecture that we created for this system, which is also described in Chapter 3, consists of four main parts.

The first part is responsible for generating the test battery of models and animations that users will assess. Through this part, the MoCap variation algorithm applies variation on the animations which are handled by the animation system. Furthermore, the simulator of movement combines the 3D models generated by the variation algorithm along with the animations to create the battery of simulations.

The second part is the analysis system and is responsible for generating the LM of the user based on their assessment values and generates a user profile that contains all the vectors that comprise the resulting LM. These vectors are the cues that the simulations were based on, the true grading of each simulation, the grade that the user has assigned to each simulation, and the resulting cue utilization as the output from the LM.

The third part is the feedback system which is responsible for generating the appropriate feedback that will be shown to the user regarding their LM such as the 3D bar charts and the text that is written on them.

Lastly, the fourth part is the training system is responsible for generating the 3D children models on top of the 3D bar charts, the animations that they are performing, as well as the piece of text explaining whether they should pay more or less attention to the valid and invalid cues.

b. What data need to be collected?

For this system to work, we tried to minimize the amount of data needed. Therefore, we only collect the three vectors that work as the input of the Lens Model (true assessment, user's assessment and cues). In addition, we store the output of the Lens Model (cue utilization) to generate the appropriate feedback and training.

RQ3: How well perceived and appreciated is the developed system?

a. Can we successfully generate a larger sample of animations by varying the invalid cues and the MoCap data?

To answer this question, we need to look into Figure 5.10 where we show the average of how many actors the participants thought we recruited to record the animations that they saw. The average response of 22 of our participants (two participants did not specify a number) was 8.65, while in reality we only recruited two actors. This shows that with only two actors, we have managed to more than quadruple the number of actors needed. This was achieved by only varying the invalid cues and applying low Perlin noise variance to the limbs of the animations. Although our method is not perfect, since the participants noticed repetitions, it is a promising approach to low-cost variation of the simulations.

b. Is the VR system and the resulting Lens Model sufficient enough to uncover the biases of participants?

Through the systematic variation we performed of valid and invalid cues on the digital children and the responses we acquired from the participants in the questionnaire and interviews, we can say that we have inconclusive results related to this question. On the other hand, we can say that our approach is a valid start for uncovering the biases since some participants acknowledged their biases that were modelled by the Lens Model based on their inputs. In general, this question will benefit from further research, where comparisons between this and traditional assessment can be performed.

c. Is the VR system sufficient enough to trigger a sensible discussion with the participants for their bias and did participants learn from it?

To answer this question, we need to look into how many participants initiated discussions about their feedback without being prompted. In total 11 participants (out of 24) discussed their feedback, by either asking questions for what it means, how they can improve on it, accepting it or denying it. This indicates the willingness of participants to discuss their outcomes and further understand their results. Even if participants deny their lens model, there is still room for discussion and training to help them realise it and understand it. This question can also benefit from further exploration from future research, although we can say that our approach is promising with regards to enabling the participants to think of and discuss their results since almost half of our sample size was willing to do so.

RQ4: How much does the VR feedback phase add to the 2D version of the feedback?

a. Do the visualizations help make the feedback phase more understandable and convincing with the outcome?

The answer to this question comes from the questions asked in the questionnaire. Participants overall rated VR as their preferred device to review the feedback on. Additionally, they rated the VR feedback as overall more valuable than the computer feedback, although the responses were close to equal. In general, they rated the VR feedback as easy to understand, with the interface decisions we made to be helpful (averages of 1.5 out of 2). For the animations on the 3D models, they gave an average of 1.25 out of 2, with the different execution qualities for the valid cues averaging on 1 out of 2. Finally, they rated that the VR feedback was more convincing than the computer feedback with 1 out of 2. Overall, it seems that the general

preference of the participants is that the VR feedback was more understandable and convincing than the computer feedback.

6.2 Discussion

The results we drew from our user study, concluded that we made a promising start in modelling the bias of users, while also attempting to educate them and kick-starting discussions about it. Participants were overall happy with the VR system, while also receiving some negative feedback. We also received some surprising comments for future improvement that will be mentioned in Chapter 6.4.

Starting with the constructive problems that were stated in Chapter 1.1, we have managed to generate the architecture of how such a system looks like, to achieve all of its goals. Through the four modules that comprise our system, we can create a test battery of simulations for users to assess, followed by the modelling of their bias using Brunswik's Lens Model. Furthermore, we can successfully create a feedback and training loop for the user that provides them with a descriptive and easily comprehensible version of their Lens Model.

As for the empirical problems of our motivation, through our user study, we have realised that the majority of our participants were very accepting towards such a system as a training tool. Some were eager to discuss their results, whereas others commented on additional features that would make the whole experience even more immersive and better suited to their personal needs as teachers or teachers-to-be. Unfortunately, we did not manage to run user studies with experts, which would have yielded even more insights as to the system's validity and reliability. Regarding modelling the bias of our participants, although we retrieve inconclusive results, it is apparent that we are headed in the right direction since almost half of our participants commented on their bias by accepting or denying it. Additionally, based on the lens model results of participants who denied their bias, we can see that they indeed rated children with, what the model indicated as their favoured, characteristics higher. Through this, we can see that there is room for improvement in terms of the feedback that we provide so that it is more convincing. This can benefit further from more variation and more targeted questions towards the bias to understand the participants' perception. Lastly, the feedback phase provided us with multiple insights that participants, in general, were happy with what we implemented in VR since they rated the VR feedback more positively than the computer feedback. This shows that we were headed in the right direction, but this can still benefit from further research and a more vast sample size of participants.

Overall, through this project, we have managed to combine our research's group study domain (Interaction Technology) with sports to create a fitting simulation for our problem. We have created a system that leverages the capabilities of state-of-the-art technologies, such as VR and MoCap, to produce a teaching mechanism. As we speculated in the introduction of this thesis report, Virtual Reality is a perfect fit for this scenario, as participants also commented on the immersiveness and the realism of the experience.

Furthermore, we have built on top of Utesch's (2020) work, and, based on our user study's results, managed to create a cost-efficient pipeline of acquiring recordings of exercises. Based on the participants' answers, we have managed to produce quadruple simulations with only two motion capture actors. One important note here is that the variation algorithm we used

for the motion capture is very simplistic and can be further improved upon. In addition, some of our participants' comments have confirmed our initial thoughts that VR is a good alternative to videos due to its immersive nature. Participants found the experience to be natural and not plastic, while also helping them be more objective in their assessment and rated the overall feedback on VR to be more than the one on the computer. More specifically, they highly rated the VR feedback phase as easy to understand and helpful, as well as the colours, text, 3D models and animations in the feedback and training phase. One drawback is the use of a yellow-coloured 3D bar chart with white text, as two of our participants stated that it was hard to read the text. This can be mitigated by changing the brightness of the colour or the colour of the text.

In addition, through the interviews, we have uncovered possible additions that would benefit the system. The most prominent comment was adding emotions to the children since the emotional drive is a major factor in the grading of a child. It shows the motivation and intention of the child which gives a holistic picture to the assessor. Moreover, having more children in the background, either performing other activities or just observing the child that is currently performing the exercises provides a more representative setting of a real classroom. In actual classrooms, children interact with one another and the simulation was lacking these interactions. Also, a baseline for how an exercise should be performed was one of the most requested additions. This can help even non-PE teachers in understanding how an exercise should be graded. These are interesting elements to research further whether they have a positive or negative impact on the system.

Concluding, this project was fruitful in terms of the final system that was created and the results from the initial evaluation. Through our results, we understand that we created an attractive and stimulating novel solution, which participants found to be consistent and easy to use. It is apparent that our participants appreciated our approach and what we set out to achieve, and provided us with valuable feedback. The project can benefit from further evaluation and work, but we made successful first steps towards an interactive teaching experience for teachers to help them improve their assessment skills.

6.3 Limitations

This thesis project, albeit fruitful in insights and answers, suffered from limitations. First and foremost, there was a notable lack of experience in 3D modelling and rigging from the researchers. This reduced the amount of variety that could be provided in different appearances in the looks of the children, as noted by some of the participants. We were also limited in the variety of hair models as well as facial characteristics.

Secondly, the limited tracking space of the MoCap equipment reduced the spectrum of exercises that could be performed and recorded only to stationary ones. This restricted the different simulations that we could show to the participants, as well as how much of the TGMD-3 framework we could cover.

Furthermore, lack of knowledge in physics simulation refrained us from implementing the ball in the exercises, which was noted and requested by multiple participants. We believe that this might have influenced the judgment of participants since the ball was an important factor in grading the children. This factor even affected the judgment of the two raters that performed the

initial assessment of the simulations to get the true grade. In addition to this, the motion capture variation algorithm we implemented was not what the background study unveiled due to time constraints and lack of knowledge with regard to animations and inverse kinematics.

A last note regarding the system is the choice of HMD. Although the choice was ideal when it comes to mobility and ease of access and distribution, the hardware of the Oculus Quest 2 is limited in performance. This restricts the amount of detail and elements that can be shown at any time, without introducing stuttering and crashes in the system.

When it comes to the study, the lack of space in the rooms that the study was conducted, restricted the participants in freedom of movement. The participants had to rely on teleporting around the environment without moving around too much since they would run out of space. Another note is that the HMD lenses were fogging up since the participants had to wear masks due to COVID-19 regulations. This irritated some participants and in some cases, they requested to take off the headset to clean the lenses and proceed with the experiment.

Last but not least, although all of the participants were English speakers and readers, their first language is German. This still introduced some barriers in terms of communication since in some cases the participants could not freely express themselves.

6.4 Future Work

As discussed in section 6.2, an important addition to the system would be to expand and improve the motion capture variation algorithm of the system to have even more variation in the animations without having to perform additional motion capture. As for the motion capture, a larger recording space or a more flexible motion capture hardware, such as the Xsens products, would be a better choice for recording the animations, so that the movement library can be expanded. Additionally, simulations of injuries, asymmetry in movement as well as irregular movement patterns based on procedural and/or physics mixed animations can be added to the animation dataset.

As for the 3D models, more detail can be added so that they represent more realistic and life-like cases of children. Facial expressions can be added to show the motivation of children, which was noted as an important factor when grading a child. In addition, the baseline of an exercise, or a bad and a perfect execution of an exercise can be shown so that users know exactly what to compare to. Lastly, more children could be added in the background to show some form of interaction between them and mimic a real-life classroom environment.

Another factor that can be further tested in the future is the combination of children and animation to generate more inputs for the Lens Model. It needs to be studied whether more inputs to the LM lead to better results for the modelling of bias.

With regards to the study, the system can be used to create a copy of the video study that Utesch (2020) conducted, in order to compare results and draw statistically significant conclusions. Such conclusions can focus on the possibly exponential expandability of the VR system with less cost when compared to the video approach.

Additionally, a larger-scale study with more participating countries can yield more universal results and more insights into how assessors in different countries perform and behave. This can generate bias models for target countries so that they can be used as examples in the target group’s education.

Last but not least, as mentioned in Section 1.2, this project can greatly benefit from a long-term study where the assessment of the teachers will be examined to see whether they have improved. This can help better solidify the contribution of this project as a teaching tool.

Concluding, our research proved to have a high potential for being useful and successful. The reaction from the participants was positive in terms of what we aimed to achieve and how we did so, with some fruitful recommendations on how to improve the system. Future work should be able to develop this system into a finalized product that can be used in real-life scenarios, with further user studies, while also providing more scientific grounding of our conclusions and limitations with more precise details.

References

Baodong, Y. (2014). Hawkeye technology using tennis match. Computer Modelling and New Technologies, 18, 400-402.

Baylor, A. L., & Kim, Y. (2004, August). Pedagogical agent design: The impact of agent realism, gender, ethnicity, and instructional role. In International conference on intelligent tutoring systems (pp. 592-603). Springer, Berlin, Heidelberg.

Baylor, A. L., & Plant, E. A. (2005). Pedagogical agents as social models for engineering: The influence of agent appearance on female choice. Artificial intelligence in education: Supporting learning through intelligent and socially informed technology, 125, 65.

Bone, A. (1999). Ensuring successful assessment (p. 32). National Centre for Legal Education.

Bonefeld, M., & Dickhäuser, O. (2018). (Biased) Grading of students' performance: Students' names, performance level, and implicit attitudes. Frontiers in psychology, 9, 481.

Bonesrønning, H. (2004). Do the teachers' grading practices affect student achievement?. Education Economics, 12(2), 151-167.

Bonesrønning, H. (2008). The effect of grading practices on gender differences in academic performance. Bulletin of Economic Research, 60(3), 245-264.

Boukhayma, A., & Boyer, E. (2017, October). Controllable variation synthesis for surface motion capture. In 2017 International Conference on 3D Vision (3DV) (pp. 309-317). IEEE.

Boukhayma, A., & Boyer, E. (2018). Surface motion capture animation synthesis. IEEE Transactions on Visualization and Computer Graphics, 25(6), 2270-2283.

Castelli, D. M., Barcelona, J. M., & Bryant, L. (2015). Contextualizing physical literacy in the school environment: The challenges. Journal of Sport and Health Science, 4(2), 156-163.

Chi, E. H. (2008). Sensors and ubiquitous computing technologies in sports. WIT Transactions on State-of-the-art in Science and Engineering, 32.

Cooksey, R. W., & Freebody, P. (1987). Cue subset contributions in the hierarchical multivariate lens model: Judgments of children's reading achievement. Organizational Behavior and Human Decision Processes, 39(1), 115-132.

Cushion, C. J., & Townsend, R. C. (2019). Technology-enhanced learning in coaching: A review of literature. Educational Review, 71(5), 631-649.

DeLuca, C. (2012). Preparing teachers for the age of accountability: Toward a framework for assessment education. Action in Teacher Education, 34(5-6), 576-591.

Douglas, D., & Gifford, R. (2001). Evaluation of the physical classroom by students and professors: A lens model approach. Educational Research, 43(3), 295-309.

Fiedler, K. (1996). Explaining and simulating judgment biases as an aggregation phenomenon in probabilistic, multiple-cue environments. Psychological Review, 103(1), 193–214.

Fitzpatrick, T. B. (1975). "Soleil et peau" [Sun and skin]. Journal de Médecine Esthétique (in French) (2): 33–34

Fuentealba, C. (2011). The role of assessment in the student learning process. Journal of veterinary medical education, 38(2), 157-162.

Greenlees, I., Buscombe, R., Thelwell, R., Holder, T., & Rimmer, M. (2005). Impact of opponents' clothing and body language on impression formation and outcome expectations. Journal of Sport and Exercise Psychology, 27(1), 39-52.

Hadjidemetriou, G., (2021). Research Topics for Interaction Technology Thesis. University of Twente.

Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. Psychological review, 71(6), 438.

Hill, M. F., Ell, F., Grudnoff, L., & Limbrick, L. (2014). Practise what you preach: Initial teacher education students learning about assessment. Assessment Matters, 7(90), e112.

Klapp, A. (2015). Does grading affect educational attainment? A longitudinal study. Assessment in Education: Principles, Policy & Practice, 22(3), 302-323.

Koch, S. C. (2004). Construction of gender: A lens-model inspired gender communication approach. Sex Roles, 51, 171-186

Lam, R. (2015). Language assessment training in Hong Kong: Implications for language assessment literacy. Language Testing, 32(2), 169-197.

Leahy, S., & Wiliam, D. (2012). From teachers to schools: scaling up professional development for formative assessment. Assessment and learning, 2, 49-71.

Lekholm, A. K., & Cliffordson, C. (2009). Effects of student characteristics on grades in compulsory school. Educational Research and Evaluation, 15(1), 1-23.

Maclellan, E. (2004). Initial knowledge states about assessment: Novice teachers' conceptualisations. Teaching and Teacher Education, 20(5), 523-535.

Malouff, J. M., & Thorsteinsson, E. B. (2016). Bias in grading: A meta-analysis of experimental research findings. Australian Journal of Education, 60(3), 245-256.

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. Educational Measurement: Issues and Practice, 20(1), 20-32.

Mertler, C. A. (2004). Secondary teachers' assessment literacy: Does classroom experience make a difference?. American secondary education, 49-64.

Mertler, C. A. (2009). Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. Improving schools, 12(2), 101-113.

Mueller, F., & Young, D. (2018). 10 Lenses to Design Sports-HCI. Foundations and Trends® in Human–Computer Interaction, 12(3), 172-237.

Oudman, S., van de Pol, J., Bakker, A., Moerbeek, M., & van Gog, T. (2018). Effects of different cue types on the accuracy of primary school teachers' judgments of students' mathematical understanding. Teaching and Teacher Education, 76, 214-226.

Pan, J., Zhang, L., Lin, M. C., & Manocha, D. (2010). A hybrid approach for simulating human motion in constrained environments. Computer Animation and Virtual Worlds, 21(3-4), 137-149.

Piek, J. P., Baynam, G. B., & Barrett, N. C. (2006). The relationship between fine and gross motor ability, self-perceptions and self-worth in children and adolescents. Human movement science, 25(1), 65-75.

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. Educational Measurement: Issues and Practice, 12(4), 10-12.

Plant, E. A., Baylor, A. L., Doerr, C. E., & Rosenberg-Kima, R. B. (2009). Changing middle-school students' attitudes and performance regarding engineering with computer-based social models. Computers & Education, 53(2), 209-215.

Poorthuis, A. M., Juvonen, J., Thomaes, S., Denissen, J. J., Orobio de Castro, B., & Van Aken, M. A. (2015). Do grades shape students' school engagement? The psychological consequences of report card grades at the beginning of secondary school. Journal of Educational Psychology, 107(3), 842.

Preuschl, E., Baca, A., Novatchkov, H., Kornfeind, P., Bichler, S., & Boecskoer, M. (2010). Mobile motion advisor—A feedback system for physical exercise in schools. Procedia Engineering, 2(2), 2741-2747.

Preuschl, E., Tampier, M., Schermer, T., & Baca, A. (2016). Introduction of the relative activity index: Towards a fair method to score school children's activity using smartphones. In Proceedings of the 10th International Symposium on Computer Science in Sports (ISCSS) (pp. 153-156). Springer, Cham.

Prøitz, T. S. (2013). Variations in grading practice–subjects matter. Education Inquiry, 4(3), 22629.

Randomness, F. (2013). Advanced Randomness Techniques for Game AI. Game AI Pro: Collected Wisdom of Game AI Professionals, 29.

Reilly, S., Barron, P., Cahill, V., Moran, K., & Haahr, M. (2009). A general-purpose taxonomy of computer-augmented sports systems. In Digital Sport for Performance Enhancement and Competitive Evolution: Intelligent Gaming Technologies (pp. 19-35). IGI Global.

Rose, B., Larkin, D., & Berger, B. G. (1997). Coordination and gender influences on the perceived competence of children. Adapted Physical Activity Quarterly, 14(3), 210-221.

Rosenberg-Kima, R. B., Baylor, A. L., Plant, E. A., & Doerr, C. E. (2008). Interface agents as social models for female students: The effects of agent visual presence and appearance on female students' attitudes and beliefs. Computers in Human Behavior, 24(6), 2741-2756.

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2014, June). Applying the user experience questionnaire (UEQ) in different evaluation scenarios. In International Conference of Design, User Experience, and Usability (pp. 383-392). Springer, Cham.

Shiban, Y., Schelhorn, I., Jobst, V., Hörnlein, A., Puppe, F., Pauli, P., & Mühlberger, A. (2015). The appearance effect: Influences of virtual agent features on performance and motivation. Computers in Human Behavior, 49, 5-11.

Sprietsma, M. (2009). Discrimination in grading? Experimental evidence from primary school. Experimental Evidence from Primary School, 09-074.

Stiggins, R. (1991). Assessment Literacy. The Phi Delta Kappan, 72(7), 534-539. Retrieved July 10, 2021, from http://www.jstor.org/stable/20404455

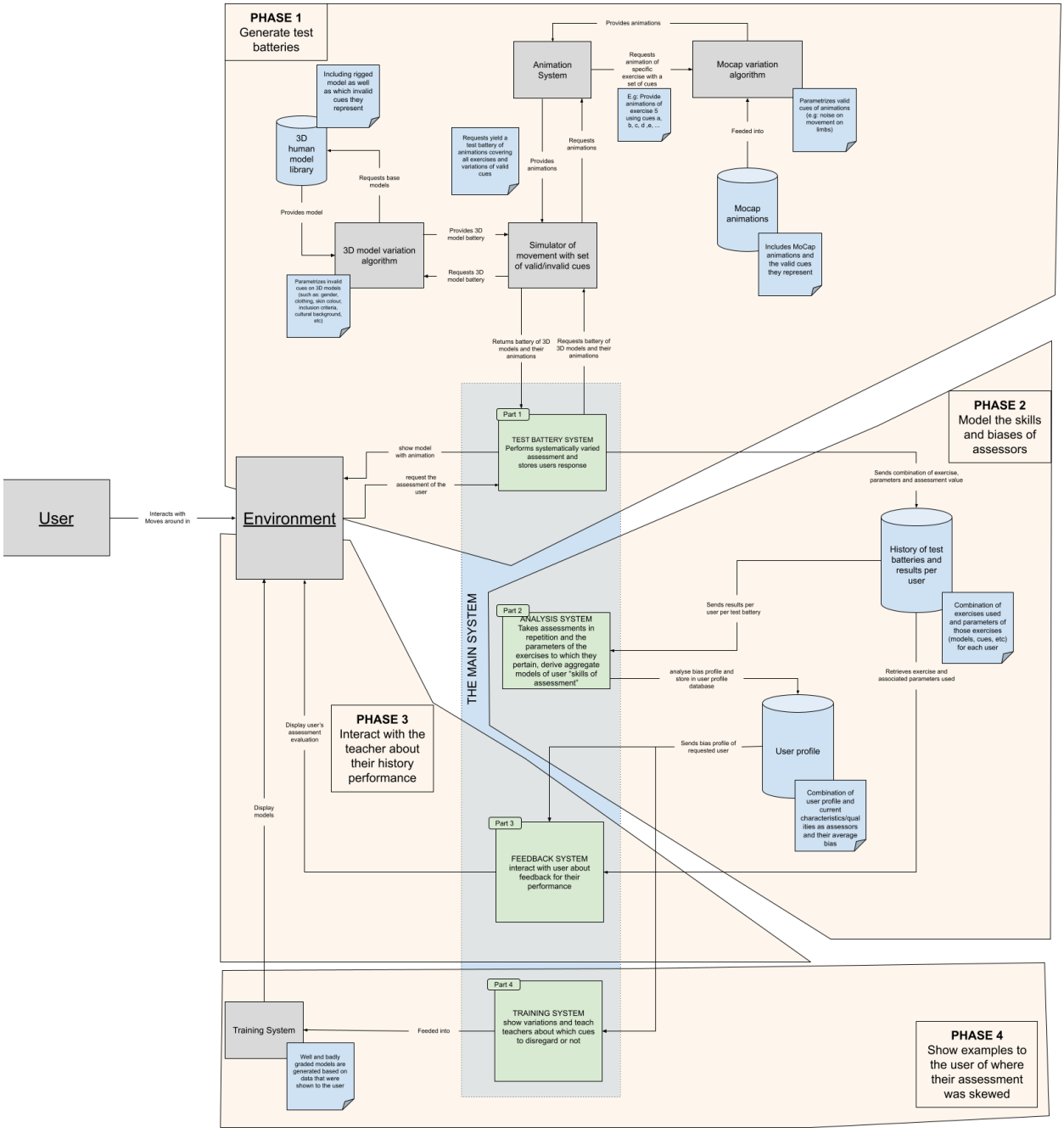
Sun, Y., & Cheng, L. (2014). Teachers' grading practices: Meaning and values assigned. Assessment in Education: Principles, Policy & Practice, 21(3), 326-343.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International journal of medical education, 2, 53.

The Standards for Teacher Competence in Educational Assessment of Students - American Federation of Teachers, National Council on Measurement in Education, & National Education Association (AFT, NCME, & NEA) Standards for teacher competence in

educational assessment of students Educational Measurement: Issues and Practice, 9 (4) (1990), pp. 30-32.
Ulrich, D. A. (1985). TGMD, test of gross motor development. Pro-Ed.
Ulrich, D.A. (2000). The Test of Gross Motor Development (2nd ed.). Austin, TX: PROED, Inc.
Ulrich, D. A. (2017). Introduction to the special section: Evaluation of the psychometric properties of the TGMD-3. Journal of Motor Learning and Development, 5(1), 1-4.
Utesch, T. (2020). A pedagogical perspective on the assessment of educational and sports data. Presented at the 1st Workshop on Action Modelling for Interaction and Analysis in Smart Sports and Physical Education (MAIStroPE), ICMI, Utrecht. Manuscript in preparation.
Williams, H. G., & Monsma, E. V. (2007). Assessment of gross motor development. Psychoeducational assessment of preschool children, 397-434.
Xu, Y., & Brown, G. T. (2016). Teacher assessment literacy in practice: A reconceptualization. Teaching and Teacher Education, 58, 149-162.

Appendix A - System Architecture



Appendix B - Questionnaire

A VR approach for modeling the assessment bias of primary school teachers and educating them about it - VR Questionnaire

Hello and thank you for taking the time to fill in this questionnaire for my Master's thesis study.

My aim with my Master's thesis, is to help inform and train teachers in becoming better assessors. They should learn which cues should and which should not be utilized when they are assessing movement skills in primary school students, in order to provide more accurate assessment for the students' skills.

In this survey you will be asked questions regarding your experience with the tool that you have used to assess the digital models of children performing a set of movement exercises in Virtual Reality. It will take approximately 15 minutes.

If you have any questions you can contact me at g.hadjidemetriou@student.utwente.nl or my supervisor at d.reidsma@utwente.nl.

If you have questions about your rights as a research participant, or wish to obtain information, ask questions, or discuss any concerns about this study with someone other than the researcher(s), please contact the Secretary of the Ethics Committee of the Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS) at the University of Twente:
Drs. P. De Willigen
ethicscommittee-cis@utwente.nl
Building: Zilverling 1051
Drienerlolaan 5, 7522 NB
Enschede, The Netherlands

* Required

1. What is your field of study? *

2. What is your level of education? *

Mark only one oval.

- ◯ Bachelor's
- ◯ Master's
- ◯ Other

3. Have you used a Virtual Reality headset before? *

Mark only one oval.

- ◯ Yes
- ◯ No

4. If yes, how often do you use a Virtual Reality headset?

Mark only one oval.

- ◯ Sometimes per year
- ◯ Sometimes per month
- ◯ Sometimes per week

System evaluation

In this section, you are asked to assess your view of the system by selecting one option for each question.

5. I found the system enjoyable to use: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

6. I found the system easy to understand: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

7. I found the system to be a creative solution: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

8. I found the system easy to learn: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

9. I found the system valuable: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

10. I found the system boring to use: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

11. I found the system interesting to use: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

12. I think the system is an inventive solution: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

13. I found the system unnecessarily complex: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

14. I found the system pleasing to use: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

15. I found the system as a usual approach to the problem: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

16. I found the experience to be pleasant: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

17. I found the system to be motivating: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

18. The system met my expectations: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

19. I found the system to be an efficient approach: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

20. I found the system unnecessarily confusing: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

21. I found the system to be practical: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

22. I think the system is well organized: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

23. I think the system is innovative: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

24. I think that I would need the support of a technical person to be able to use this system: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

25. I thought there was too much inconsistency in this system: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

26. I found the system to be repetitive: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

27. I found the animation control tools provided to be sufficient: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

28. I believe that the goal of the system was clearly communicated: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

Feedback phase evaluation

In this section, you will be asked questions regarding your experience with the feedback phase of the system.

29. I found the feedback phase in VR to be valuable: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

30. I found the feedback phase on the computer to be valuable: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

31. Do you think the purpose of the feedback phase on the two devices was different? If yes, how so? *

32. I found the feedback phase in VR easy to understand: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

33. I think the colours of the pillars of the feedback phase in VR helped me understand the feedback: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

34. I think the addition of the 3D models in VR helped me understand the feedback: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

35. I think the addition of the animations of the 3D models in VR helped me understand the feedback: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

36. I think the addition of the different execution quality animations in VR, for the valid cues, helped me understand the feedback: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

37. I believe that the feedback phase on VR convinced me more on my assessment skills than the feedback phase on the computer: *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

38. I prefer the feedback phase on the *

Mark only one oval.

- computer
- VR headset

39. Based on your answer above, please specify why: *

40. (Optional) Were there any elements on the other device that you prefere more? If yes, what are they?

Appendix C - Interview

1. What did you like in the experience? (Why/Elaborate)
2. What did you not like in the experience? (Why/Elaborate)
3. What would you want to see in the experience? (Why/Elaborate)
4. What would you want to be removed from the experience? (Why/Elaborate)
5. Did you notice anything regarding the animations?
6. To transfer the animations into the game, we motion captured people performing these exercises. The animations of how many different people do you believe you saw?
7. Have you seen an animation more than once?

Appendix D - Informed Consent Form

Informed consent form for “A VR approach for modeling the assessment bias of primary school teachers and educating them about it” - VR study

Dear participant,

Thank you for considering participating in my study.

My aim with this project is to help inform and train teachers in becoming better assessors. They should learn which cues should and which should not be utilized when they are assessing movement skills in primary school students in order to provide more accurate assessment for the students' skills.

You can participate if you are:

- Competent adult
- teacher or teacher-in-training (PE or non-PE)
- English reader and speaker

The study will consist of using a Virtual Reality environment through the Oculus Quest 2 device as well as a laptop device. In the Virtual Reality environment you will be asked to assess the performance of a variety of 3D models of digital students performing exercises based. You will be asked to move around the environment, assess the 3D models using the tools supplied in the system and review the feedback that we will provide in Virtual Reality. On the laptop device you will be asked to review your feedback. Following this, a short interview will be conducted where you will be asked a series of questions regarding your experience followed by a questionnaire that will target more specific questions about your experience. The entire study will take approximately 1 hour and 30 minutes to complete.

Through the use of the Virtual Reality system it is possible that you will incur motion sickness, which is a physical discomfort (dizziness and/or nausea) that occurs when you move in Virtual Reality, while your physical body is standing still, and therefore your brain receives conflicting signals about movement. If this occurs, you have the freedom to stop the study. In addition, through this study we might uncover biases in specific areas related to how you assess students, such as gender, skin colour, cultural background, disability and economical status.

The gathered data will be your responses to the interview and the questionnaire, your assessment values to each of the 3D models, and a video recording of your body movements while interacting with the Virtual Reality system. All the data gathered will be anonymous, and no one will be able to link your responses back to you. Furthermore, the data will not be shared with a third party. Lastly, information collected for this study will be used for writing my master’s thesis report and for possibly writing a scientific paper.

Your participation remains at all times voluntary and you may stop the study at any point without giving any reasons. You may also request within 24 hours for your data to be deleted and/or excluded from the research by contacting the researcher of this study. Your data will be stored in a secure location for the duration of two (2) years.

Study contact details for further information:

Giorgos Hadjidemetriou

g.hadjidemetriou@student.utwente.nl

Brammelerstraat 15, 7511JG

Enschede, The Netherlands

Contact Information for Questions about Your Rights as a Research Participant

If you have questions about your rights as a research participant, or wish to obtain information, ask questions, or discuss any concerns about this study with someone other than the researcher(s), please contact the Secretary of the Ethics Committee of the Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS) at the University of Twente:

Drs. P. De Willigen
ethicscommittee-cis@utwente.nl
Building: Zilverling 1051
Drienerlolaan 5, 7522 NB
Enschede, The Netherlands

Participant #____

Consent Form for "A VR approach for modeling the assessment bias of primary school teachers and educating them about it"

YOU WILL BE GIVEN A COPY OF THIS INFORMED CONSENT FORM

Please tick the appropriate boxes	**Yes**	**No**
Taking part in the study		
I have read and understood the attached study information, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	☐	☐
I consent voluntarily to be a participant in this study.	☐	☐
I understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	☐	☐
I understand that taking part in the study involves an interview which will be recorded in the form of written notes, a questionnaire which I will need to fill in and a video recording of my body movements while using the Virtual Reality system.	☐	☐
I understand that taking part in the study may result in uncovering biases that I may have in specific areas related to how I assess students, such as gender, skin colour, cultural background, disability and economical status.	☐	☐
Risks associated with participating in the study		
I understand that taking part in the study involves the risk of motion sickness (as described in the information brochure)	☐	☐
Use of the information in the study		
I understand that information I provide will be used for writing a master's thesis report and potentially a scientific paper.	☐	☐

I understand that personal information collected about me that can identify me, such as my answers to the interview, or questionnaire and the video recordings of my body movements while using the Virtual Reality system, will not be shared beyond the study team. ☐ ☐

I agree that my information can be anonymously quoted in research outputs ☐ ☐

I agree to be audio/video recorded. ☐ ☐

Signature

_____	_____	_____
Name of participant	Signature	Date