sustainability

MDPI

Article

Evidence-Based Assessment of Student Performance in Virtual Worlds

Manuel Palomo-Duarte [1], Anke Berns [2], Antonio Balderas [1,*], Juan Manuel Dodero [1] and David Camacho [3]

[1] Departmento de Ingeniería Informática, Universidad de Cádiz, 11519 Puerto Real, Spain; manuel.palomo@uca.es (M.P.-D.); juanma.dodero@uca.es (J.M.D.)
[2] Departamento de Filología Francesa e Inglesa, Universidad de Cádiz, 11003 Cádiz, Spain; anke.berns@uca.es
[3] Departamento de Sistemas Informáticos, Universidad Politécnica de Madrid, 28031 Madrid, Spain; david.camacho@upm.es
* Correspondence: antonio.balderas@uca.es

Abstract: Virtual Worlds (VWs) are popular tools for teaching/learning in the twenty-first century classroom. The challenge remains however, to provide the means by which teachers could sustainably analyse and assess the performance of large groups of students in such environments. Unfortunately, external game features such as game scores and play duration have turned out to be unfair in some assessments. In this context, a case study was carried out in a foreign language course, illustrating how teachers could easily retrieve a number of performance indicators from VW-interaction logs and harness them to conduct a fine-grained analysis of students' performance, while facilitating at the same time valuable tools for their assessment. Objective performance indicators in a server database were made accessible using an end-user development programming language. This way, a range of data visualisation methods could be employed to contrast different assumptions regarding learner performance when playing a VW-based game, which was designed to help CEFR A1 level students to learn German. This way, factors such as randomisation of game tasks, which could negatively affect learner performance, were alleviated.

Keywords: virtual worlds; VW-based games; end-user development; programming language; interaction logs; assessment; language learning

Citation: Palomo-Duarte, M.; Berns, A.; Balderas, A.; Dodero, J.M.; Camacho, D. Evidence-Based Assessment of Student Performance in Virtual Worlds. *Sustainability* **2021**, *13*, 244. https://doi.org/10.3390/su13010244

Received: 30 November 2020
Accepted: 23 December 2020
Published: 29 December 2020

1. Introduction

The use of Virtual Worlds (VWs) and VW-based games in education is not new. This is evidenced by the large body of literature on VW-based teaching and learning experiences [1–3]. However, despite extensive data indicating that VWs are effective teaching/learning tools, the use of VWs for educational purposes has not been as widespread as might be expected [4]. This may be due to high development costs and/or technological barriers, administrative hurdles as well as the inherent complexities of monitoring and analysing the learning process, specially when the number of students increases [5,6].

The idea of using VW-based games for learning is twofold: on the one hand, to make the learning process more enjoyable and on the other, to provide learners with interactive environments which offer them opportunities to practice and foster different skills from their respective area of study. In terms of teaching, the main challenge remains however, to provide the means by which teachers could sustainably analyse students' interaction with the game-environment and use this information for assessing students' performance in the targeted skills [7]. In this context, the current study aims to give answer to the following research question: Can VW interaction logs help teachers to sustainably conduct a fine-grained analysis of students' performance and facilitate their assessment in VW-based game environments?

The study is based on the assumption that the information retrieved from VW interaction logs could help analysing learner performance and offer teachers, even with few programming skills and large groups of students, valuable information on students' performance. This paper illustrates how teachers can easily retrieve refined information from interaction logs by employing an end-user development (EUD) programming language and use that visual information to harness teaching and learning as well as assessment processes. To this end, a case study was conducted in a CEFR A1 level German language course during which students played a VW-based game.

2. State of the Art

In recent years many experts have recognised the enormous educational and motivational potential of video games and VWs [8–11] yet, to date, there are few empirical studies directly exploring their impact on learning processes [12–14].

The interaction that takes place in VWs aimed at learning a target language, both among students as well as between students and virtual agents, favours the development of reading and discourse management skills [15]. Peterson applied *massively multiplayer online role-playing games (MMORPGs)* [15] and *3D multiuser virtual environment (MUVE) Second Life* [16] respectively for language learning. Students' feedback suggested that the benefits of using these tools were valuable opportunities to practice the target language, learning new vocabulary, while at the same time enjoying the learning process.

Both off-the-shelf and tailor-made games, and most VWs, are restricted by private software licences; hence, in general, do not allow teachers to create nor tailor them to specific learning needs and curricular requirements, nor to trace and analyse learner interaction [5,17].

Despite the widespread trend to use external learning assessment procedures [7,18] to measure the learning impact of game-based learning experiences, conventional (pre- and post-tests) as well as other more recent completion criteria (e.g., teacher observation, the game levels played), entail some important limitations since information about the game experience itself is barely considered [12]. However, in-game assessment can provide more detailed information for analysing and assessing students' learning process, even though it is more difficult to implement by non-technical staff [19].

Some previous studies [20,21] underlined the need for a deeper analysis of learner interaction with a view to streamline the assessment of learner performance. In a research study presented by Hsiao, Lan, Kao and Li [22], student records were analysed to examine the relationship between learning outcomes and learning paths and strategies within a VW for learning Mandarin Chinese. The analysis was done by using *Second Life learning latabase (SLLDB)*, a computer tool that was developed by the authors for the purposes of their research study to record students' interaction. Although this proposal does not allow teachers to adapt the analysis of students' interaction to the specific needs of their subject, it does provide an initial effort to analyse large amounts of language learning data and its relationships with the learning outcomes. Other works [23] have studied the visualisation of students' progress when learning through game-based environments as a way of tracking students' interaction and measuring their activity in VW-based learning environments. Students' interaction is depicted by means of online graphs, illustrating their activity based on indicators such as voice and text chat communications as well as user sessions. Beyond regular graphs, Minović, Milovanović, Šošević and González [24] have used coloured circular views to visualise time-related data describing students' progress on a number of key concepts. Other types of circular views, such as semantic spiral timelines, have been used to explore the use of virtual learning environments across time [25]. Moreover, some authors have proposed the use of dendrogram representations which show how VW student communities can be visually identified after clustering avatar positions [26].

While the previously mentioned studies provided computer experts with some valuable techniques for analysing and assessing students' in-game performance, the current

work intends to make in-game assessment available to a broader audience with no need for having specific programming skills by using an EUD programming language. The EUD programming language will be developed as a Domain-Specific Language (DSL). A DSL is a programming language or specification designed to solve a specific problem. Its use is spreading, as it provides users with a programming language that is customised to the user's domain [17,27]. To illustrate how an EUD programming language can be used in the area of foreign language learning, the authors have carried out a study with a group of undergraduate students from a German language course playing a VW-based game.

3. Methodology

For this study, the authors have used the design and creation research strategy defined by Oates [28] which focuses on developing information technology (IT) artefacts such as constructs, models, methods, etc. in order to solve complex problems. In order to answer our initial assumption that accessing students' interaction logs, when playing VW-based games, could help teachers (in this case foreign language teachers) conduct a customised fine-grained analysis of students' performance, we here propose an evidenced-based method. This method aims to provide teachers with the means by which they could easily access and analyse students' interaction logs in a sustainable way (that is, independently from the number of students to be assessed).

3.1. Evidence-Based Method

In this paper, we propose the use of a method that has previously been applied in other virtual environments to assess students' performance through their interaction with both, a wiki environment [29] and a learning management system [30,31].

The evidence-based method implies working in different refinement phases until an objective and valid indicator is found that could help teachers assessing their students' performance when completing a given learning task. Next, the proposed method and the different phases that need to be taken are described in detail:

1. *Instruction*: first students need to receive instructions on how to play the game in order to complete the given learning task.
2. *Task performance*: next, students are asked to play the game.
3. *Assessment*: in order to monitor and assess students' performance, after or during the game, the teacher needs to retrieve different indicators from the VW regarding concrete student skills. These indicators can be retrieved from the VW interaction logs and later be refined with the help of the following evidence-based method (Figure 1):
    (a) *First assessment proposal*: the teacher defines an initial proposal to assess students' performance with regard to a specific skill, using for this purpose the information stored in the VW interaction logs.
    (b) *Submission*: the assessment proposal is coded in a Virtual World Query Language (VWQL) query. A VWQL query specifies a requests of concrete information from the VW interaction logs. After being coded, the query must be submitted to EvalSim, the system that interprets VWQL querys [20].
    (c) *Data collection*: the system provides the data requested from the logs.
    (d) *Results*: the system delivers several reports and figures with the data requested.
    (e) *Analysis*: the teacher analyses the data in order to decide whether they are valid indicators to assess the targeted skill/s.
    (f) *Validated*: the process ends once the data obtained are valid indicators for assessing the targeted skill/s.
    (g) *Refinement*: in case the data are not valid or need to be refined, the entire process is repeated, and a new query is launched.

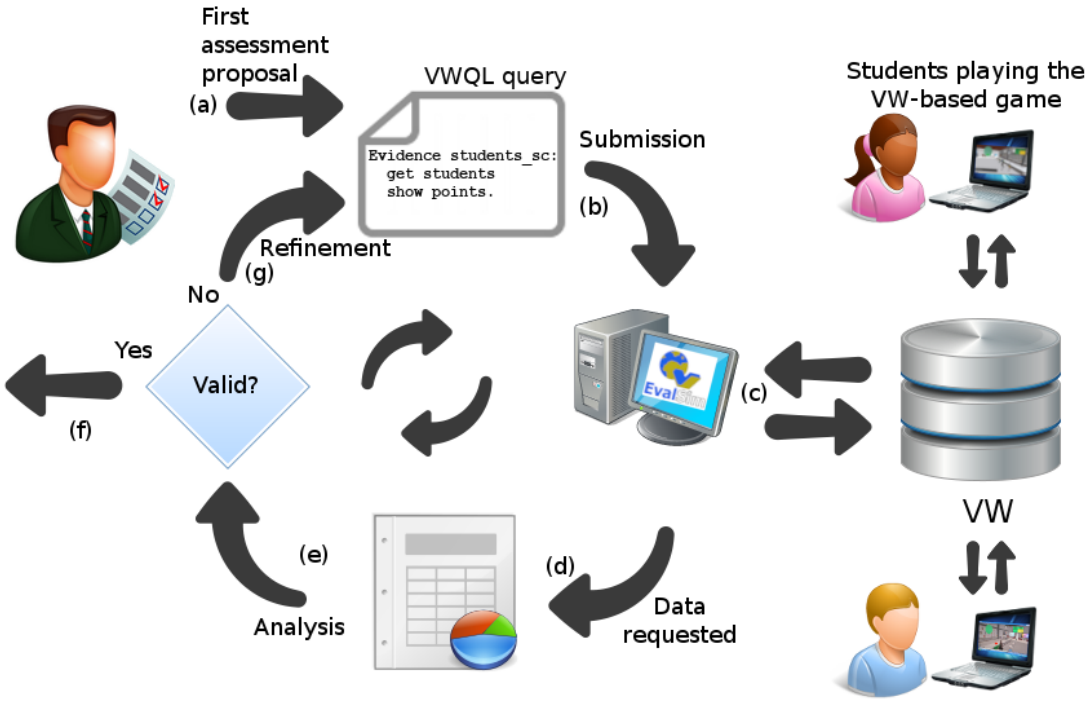

Figure 1. Scheme of the evidence-based method.

3.2. Implementation

To implement the evidence-based method two IT artefacts are needed: firstly, a VW-based game in which students are required to interact while developing the targeted skill/s and secondly, an EUD programming language, implemented as a DSL, to provide teachers with a language that helps them with coding the assessment proposals.

3.2.1. Virtual World

For the current study, a tailor-made competitive two player VW-based game called GEFE (German Expert and Fast waitEr), designed in line with specific learning needs and curricular requirements for CEFR A1-level German language learners, was used. Since some of the key items of students' language curriculum are the learning of basic vocabulary and structures that are for instance needed to perform daily tasks in the target language (e.g., ordering in a restaurant or cafeteria, buying in a supermarket, etc.), the focus here was on providing students with the opportunity to practice those items by means of a competitive role-play game. The game recreates an outdoor cafeteria scenario, using simple artificial intelligence algorithms to randomly place a total number of 12 virtual client bots all over the playing field. Additionally, the algorithm assigns to each player one waiter avatar. Play starts with the two waiter avatars lined up along the counter. Each waiter avatar will receive a call for service from one of the client bots. To attend a client, the waiter avatar must first approach the table of the respective client bot and then ask him/her for the items he/she wants to order. Interaction between both the waiter avatar and client bot takes place via text chat and in the target language. Once the order has been taken, the waiter avatar must go back to the counter and prepare the order (each order consists of 3 items: a beverage, a dish and a dessert) by identifying and clicking on the image of the different items. Hereafter the waiter avatar must return to the client's table and deliver the order (Figures 2 and 3).

Figure 2. The waiter avatar preparing a client's order.

Figure 3. The waiter avatar delivering the order to a client.

Once the waiter avatar has delivered the order, the student is provided with immediate feedback in the form of points. Points are calculated as follows: for each item which has been correctly delivered from the order, the student is awarded 1 point with 3 points being the maximum for each client. Once a waiter avatar has delivered the order, he/she receives a new call for service from another client bot. This way, if a student makes a mistake when delivering items, the maximum number of points he/she can get in the game is reduced. Although the server initially assigns the same number of client bots (12 bots) to each waiter avatar, the play round finishes after one of the two waiter avatars has attended all client bots assigned to him/her.

Due to random task assignment each game session differs from the previous one. The randomness of client bots and the distance that client bots are placed over the playing field means that the time spent to complete a required game task varies in each game session.

In order to ensure that student logs are stored correctly, regardless of possible drawbacks due to the server and internet connection, player movement traces are registered throughout the game and by means of player coordinates on a regular basis.

3.2.2. Domain Specific Language

A Domain Specific Language (DSL) is a programming language which allows an expert to formalise and represent specific knowledge on a particular field or topic, which in this case is foreign language learning assessment.

To code the assessment proposals, the DSL used for the analysis was VWQL (available at https://bitbucket.org/RaulGS/vwql/). VWQL was developed by the authors in the context of previous works in order to retrieve information on German language learners' performance by analysing their interaction when learning through VW-based game environments, developed with OpenSim [20]. VWQL queries are submitted to EvalSim, which processes them and generates reports with the requested data. The technical development documentation and the user manual are available in [32].

The information made available by using VWQL were firstly, words and sentences (complex sentences and one-word sentences) employed by the learners while interacting

via text-chat, secondly, the turns and time students needed to accomplish the given game task and thirdly, the number of clients attended, and points obtained. Below, we can see the reserved words and syntax of VWQL.

```
Evidence name_of_the_evidence:
get students [student_id]
show (words [dict] | sentences | single | turns | time | clients
| points) +.
```

Moreover, the use of VWQL allows to obtain two types of information: a VW tracking map for each game session and a general data report. A VW tracking map is a visualisation method that allows teachers to visually analyse students' movements, when playing VW-based games. This process is complemented by a general data report i.e., a set of files containing students' movement traces and interaction logs. Additionally, a spreadsheet program can be used to process files, analyse data and generate charts and graphs.

4. Evaluation

The evaluation carried out for the current study was done with 16 undergraduate students from a CEFR A1-level German language course and in collaboration with the students' language teacher being one of the authors of the study. All participants had previously been enrolled in a 2-semester German language course at a Spanish university (6 ECTS/semester) for a total of 96 classroom contact hours and 204 independent learning hours.

The experiment was organised in two sessions: the first session aimed to familiarise students with the VW-based game environment; the second required students to play the game by means of waiter avatars and to interact with a number of client-bots using the target language (German) as the only vehicle for communication. Each game session lasted a maximum of 25 minutes. Table 1 shows the different game sessions and students who participated in each session.

Table 1. List of games sessions and their respective participants.

Game Session	Students
1	Stud1 and Stud2
2	Stud3 and Stud4
3	Stud5 and Stud6
4	Stud7 and Stud8
5	Stud9 and Stud10
6	Stud11 and Stud12
7	Stud13 and Stud14
8	Stud15 and Stud16

With a view to analyse and assess students' language performance when playing the game, the course teacher intended to look at two aspects: students' vocabulary knowledge and students' interaction in the target language.

4.1. Vocabulary Knowledge

In order to assess students' vocabulary knowledge, the teacher first analysed their game performance in terms of game scores. This procedure was based on the language teacher's assumption that the more clients a student has attended and orders he/she was able to successfully deliver, the better was his/her vocabulary knowledge. Consequently, the more items he/she was able to deliver the more game scores he/she obtained.

4.2. Interaction in the Target Language

Second, the teacher assessed students' interaction by analysing the time a student spent to interact with the client bots, to perform the given game task, and the game scores (points) he/she obtained. This procedure was based on the assumption that the less time a student needed to perform the given task correctly, the better was his/her ability to interact and negotiate effectively in the target language.

5. Results

The purpose of the following section is twofold: firstly, to show the assessment phase (Figure 1), that was carried out by the teacher in order to evaluate the aforementioned language skills, and to discuss the suitability of the evidence-based method to help teachers assessing different fine-grained aspects of students' language learning process.

5.1. Vocabulary Knowledge: Analysis and Validation

In the following, the steps taken to assess students' vocabulary knowledge are described.

5.1.1. First Assessment Proposal

Students' vocabulary knowledge was analysed by their game performance in terms of game scores.

- *Submission*: the assessment proposal is coded through the VWQL query shown below.

```
Evidence students_scores:
get students
show points.
```

- *Results*: the system delivers various reports and figures with the data requested. Figure 4 illustrates the game scores (points) obtained by each student. Game scores range from 0 to 34, with a standard deviation (SD) of more than 10.

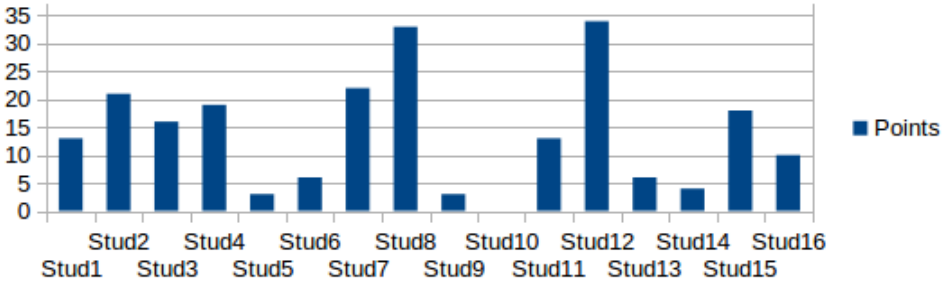

Figure 4. Student game scores (points).

- *Analysis*: regarding the assessment proposal, the data illustrate on the one hand, that two students (Stud8 and Stud12) performed significantly better than the rest of the participating students (Stud8 obtained 32 and Stud12 obtained 33 points, while the rest of students obtained 22 or less points). And on the other hand, that six (37.5%) of the participating students (Stud5, Stud6, Stud9, Stud10, Stud13, Stud14) only obtained 6 or less points, which means a score that was five times lower than the score obtained by Stud8 and Stud12.
- *Validated*: although the obtained report provides valuable information to help teachers assessing students' performance in terms of vocabulary knowledge, the teacher became aware that she needed more detailed information on students' performance in order to validate the retrieved assessment indicator. For instance, in case a student failed several client-orders, but attended much more clients than another student, the first one would probably obtain a higher score than the second one, who delivered his/her orders more precisely. Thus, the proposed indicator should be refined by also considering the number of orders that have been delivered incorrectly.

- *Refinement*: to refine the game score indicator the teacher will need to analyse the number of orders each student delivered successfully. To obtain this new indicator, the teacher will need the points obtained by each student as well as the number of clients attended.

5.1.2. Second Assessment Proposal

Students' vocabulary knowledge will be analysed by their game performance in terms of points/clients.

- *Submission*: the assessment proposal is coded through the VWQL query shown below.

```
Evidence students_points_per_clients:
get students
show points, clients.
```

- *Results*: The values obtained for the points/client ratio range from 0 points when the student failed to deliver any of the items in the order to a maximum of 3 points when the student successfully delivered all 3 items to the client (Table 2).

Table 2. Game score, number of clients and points/client ratio values per student.

Student	Game Score	Number of Clients	Points/Clients Ratio (Score per Client)
Stud1	13	5	2.60
Stud2	21	7	3.00
Stud3	16	8	2.00
Stud4	19	8	2.38
Stud5	3	1	3.00
Stud6	6	2	3.00
Stud7	22	9	2.44
Stud8	33	11	3.00
Stud9	3	1	3.00
Stud10	0	1	0.00
Stud11	13	5	2.60
Stud12	34	12	2.83
Stud13	6	2	3.00
Stud14	4	2	2.00
Stud15	18	6	3.00
Stud16	10	5	2.00

- *Analysis*: with regard to the assessment proposal, the data from Table 2 illustrate that three students (Stud2, Stud8 and Stud15) completed the given game task very precisely, delivering correctly all their clients' orders. Nevertheless, student scores depend not only on the success of each student's deliveries, but also on the time they played, which could differ in each game session. Therefore, it is not only the score that can be considered, but also the number of clients that the student was able to serve, which were necessarily related with the time the student played. The authors assume that some of the students might be more reflective learners and thus would have needed more time to perform the same game task and to obtain a higher game score. Finally, to determine whether a student had accomplished the game task satisfactorily, the teacher could set a score threshold, by establishing the points (ratio per order) a student would need to be considered as a learner with a strong vocabulary knowledge. Such a score threshold could establish, for instance, that obtaining 2.60 (see Stud1 and Stud11) or more points (see Stud12) would be an indicator for having a strong vocabulary knowledge, regardless the fact that Stud1 and Stud11 attended less than half of the clients Stud12 was able to attend in the same time. Although four students (Stud5, Stud6, Stud9 and Stud13) also completed all of their deliveries accurately, they are considered outliers for having served two clients or less.

- *Validated*: from the teacher's point of view, the information on the points/client ratio provides a valid indicator with regard to students' vocabulary knowledge, since she was able to determine whether her students had successfully acquired the targeted vocabulary from the course syllabus. In fact, the data show that all learners, except one (Stud10), obtained a good points/client ratio (in the range of 2 to 3).
  Although the teacher used the point per order indicator to source a more precise assessment of students' vocabulary knowledge, she considered it necessary to additionally establish a minimum number of clients attended. This was based on analysing the performance of Stud5, Stud6, Stud9, Stud10, Stud13 and Stud14, since these students attended 2 or fewer clients. The data therefore suggest that the validity of the point per order indicator might not be sufficient for the assessment of the respective students, since they attended less than 17 percent of the total amount of clients.

5.2. Interaction in the Target Language: Analysis and Validation

In the following, the authors describe the steps that were taken to assess students' interaction skills in the target language.

5.2.1. First Assessment Proposal

Students' interaction in the target language will be assessed by analysing the relationship between students' game scores and playing time. This means, that the average score per minute (points/minute ratio) could help assessing students' performance in terms of effectiveness, providing some valuable indicator for evaluating students' ability to interact in the target language.

- *Submission*: the assessment proposal is coded through the VWQL query shown below.

```
Evidence students_points_per_minute:
get students
show points, time.
```

- *Results*: the system delivers several reports and figures, providing detailed information on the game scores and play duration of each student. Since play duration is shown in minutes, a third indicator, named points/min ratio, was added by using a spreadsheet software (Table 3). The results from Table 3 indicate that values for the points/minute ratio range from 0.50 to 3.68.

Table 3. Game score play duration and points/minute ratio values per student.

Student	Game Score	Play Duration (in Minutes)	Points/Minute Ratio
Stud1	13	9	1.44
Stud2	21	9	2.33
Stud3	16	9	1.78
Stud4	19	9	2.11
Stud5	3	2	1.50
Stud6	6	2	3.00
Stud7	22	9	2.44
Stud8	33	9	3.67
Stud9	3	3	1.00
Stud10	0	3	0.00
Stud11	13	9	1.44
Stud12	34	9	3.78
Stud13	6	8	0.75
Stud14	4	8	0.50
Stud15	18	7	2.57
Stud16	10	7	1.42

- *Analysis*: with regard to the assessment proposal, the data illustrate that two of the participating students (Stud12 and Stud8) performed significantly better than the rest of the students: while Stud12 obtained 3.78 points/minute ratio, Stud8 obtained 3.67 points. Additionally, a look at students' game scores highlight that Stud8 and Stud12 are also the students with the best results. In order to compare both indicators, a ranking of students based on points/minute and game score is shown in Table 4. A comparison of both indicators shows that only in the case of Stud6 the assessment could significantly differ, when considering one or another indicator. In fact, Stud6 obtained very few points (6 points) since they played for less time than the rest of the top ranked students (Table 3). Nonetheless, Stud6 performed well in the given game task, delivering all client orders correctly, which explains why Stud6 obtained a relatively high game score, obtaining similar results compared to the top ranked students (Stud12, Stud8, Stud7, Stud2 and Stud4).
  However, the points obtained by students who played during significantly less time than the rest of the group had to be carefully considered: while Stud6 ratio was extremely high, those of Stud9 and Stud10 were really poor, and only Stud5 obtained an average result. Therefore, such measures need further contrast to be considered in the same terms that those of the rest of the players.

Table 4. Ranking comparison between points/minute and game scores.

Student	RANKING Position Based on Points/Minute	RANKING Position Based on Game Score
Stud12	1st	1st
Stud8	2nd	2nd
Stud7	3rd	3rd
Stud6	3rd	11th
Stud2	5th	4th
Stud4	6th	5th
Stud15	7th	6th
Stud3	8th	7th
Stud5	9th	14th
Stud11	10th	8th
Stud1	10th	8th
Stud16	12th	10th
Stud9	13th	14th
Stud13	14th	11th
Stud14	15th	13th
Stud10	16th	16th

- *Validated*: from the teacher's point of view, the information on the points per minute ratio provides a valid indicator for assessing students' ability to interact in the target language, since it allows to gather information on students' efficiency when performing the given game task. This means, given a minimum of 4 min played, the points per minute ratio indicator will show the time each student needed to correctly deliver his/her clients' orders.
- *Refinement*: despite accepting the validity of the points per minute ratio indicator, the teacher decided to consider another external factor, which could have influenced students' game performance and therefore explain some of the observed differences between student players, in terms of efficiency. Due to random task assignment each game session differed from the previous one hence requiring from each player a different effort in order to attend client orders and to complete the given game task successfully.

5.2.2. Second Assessment Proposal

Students' interaction in the target language will be analysed by considering the potential relation between students' point per minute ratio and the effort required to accomplish the given task. For the purposes of this study, effort rather than referring to student initiative is synonymous with the distance a student had to cover in order to attend his/her client orders. Note that clients are randomly placed by the system and thus, each student should cover a different distance to attend the same number of clients.

- *Submission*: the assessment proposal is coded through the VWQL query shown below.

```
Evidence students_distance_clients_points:
get students
show distance, clients, points.
```

- *Results*: the system delivers several reports and figures, providing detailed information on the distance covered by each student player pair. A look at Figures 5–7 shows how the retrieved information can be illustrated by means of a VW tracking map, providing some interesting insights into students' performance, based on student players' movement and the distance covered.

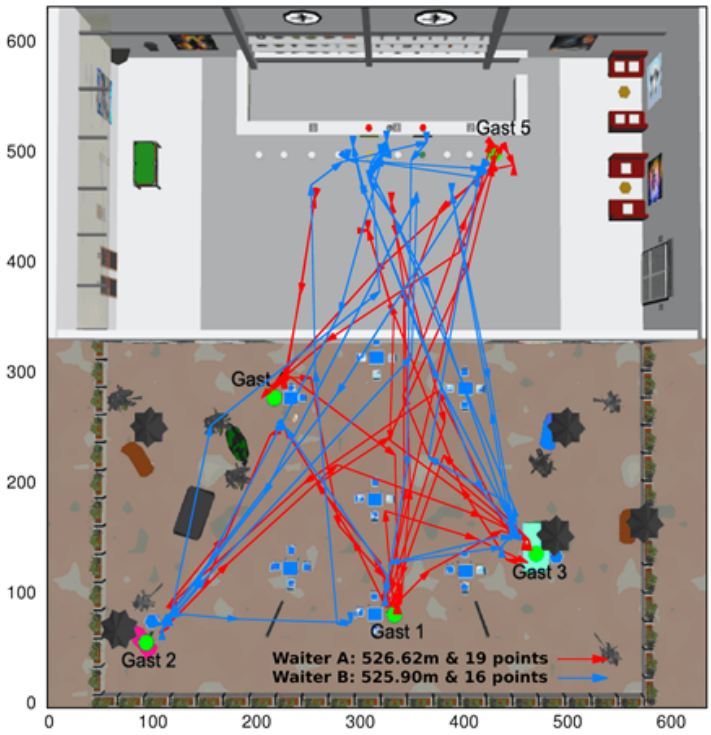

Figure 5. VW tracking map with similar distance between waiters.

- *Analysis*: a look at the different VW tracking maps (Figures 5–7) reveals that not all students were required to cover the same distance in order to attend to their clients. This implies that some students must make more of an effort to complete the same game task than others. Additionally, in terms of assessment, it means that conventional performance assessment criteria such as game score alone are clearly insufficient.

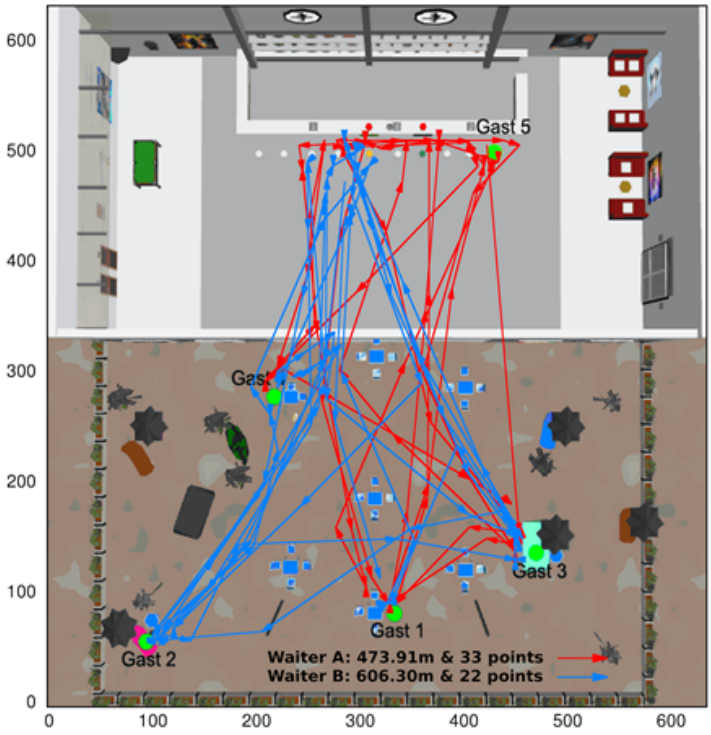

Figure 6. VW tracking map with considerable distances between waiters.

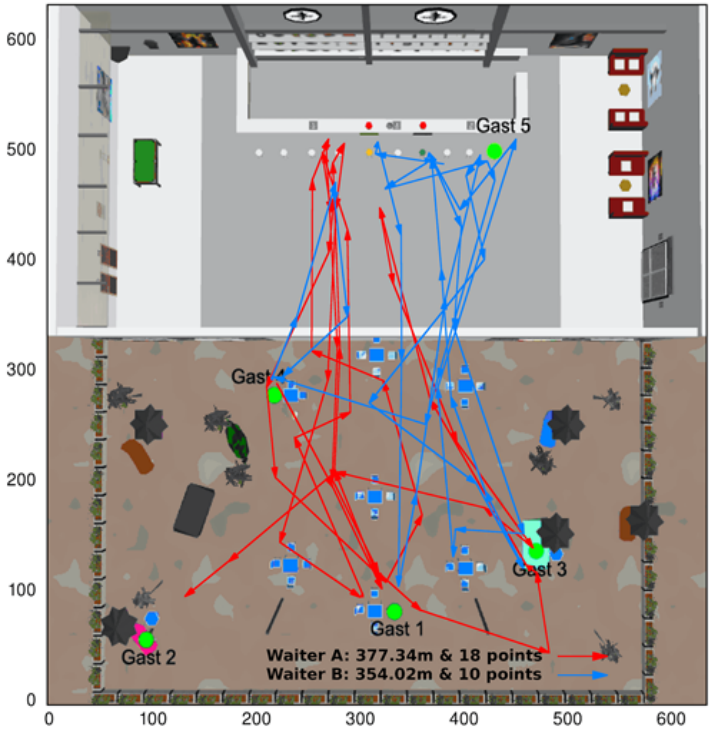

Figure 7. VW tracking map for Stud15 and Stud16.

For instance, the data in Figure 5 which show the interaction and distance covered by Stud4 (Waiter A) and Stud3 (Waiter B) clearly indicates that both students faced a similar challenge in terms of distance and hence effort required to deliver their clients' orders. In this case, the student who performed best (Stud4), having a higher game score showed clearly a better knowledge of the target language compared to the one who

performed worse (Stud3). Nonetheless, a look at Figure 6 illustrates that in the game session played by Stud7 and Stud8, Stud8 (Waiter B) needed to cover a much longer distance (146.08 metres and thus 30 per cent longer) compared to Stud7 (Waiter A) in order to attend two clients less. In fact, while Stud 7 (Waiter A) attended 11 clients, Stud8 (Waiter B) attended only 9 clients. Hence, considering game score alone would not be fair to assess students' performance.

A third case is finally illustrated in Figure 7. The data from Figure 7 show that Stud15 (Waiter A) attended, in the same amount of time, one more client than Stud16 (Waiter B) hence being more efficient. Regardless of the fact that Stud15 (Waiter A) needed to cover a greater distance and thus make a greater effort, he/she was able to attend much better to his/her clients' orders than Stud 16 (Waiter B). A look at students' game scores confirms the authors' assumption, clearly indicating that Stud15 has a stronger knowledge of the target language compared to Stud16: while Stud15 (WaiterA) obtained 18 points (a ratio of 100% regarding the delivered items), Stud16 (WaiterB) obtained only 10 points (resulting from an average ratio of 66%).

After analysing the different VW-tracking maps, we tried to identify a general group behaviour based on the data normalisation (Figure 8). While the blue line indicates the normalised clients (i.e., the clients attended by the students divided into the maximum number of clients attended by a student in the group), the red line indicates the normalised distance (i.e., the metres run by the students divided by the maximum metres run by a student in the group). Students are sorted according to the normalised clients.

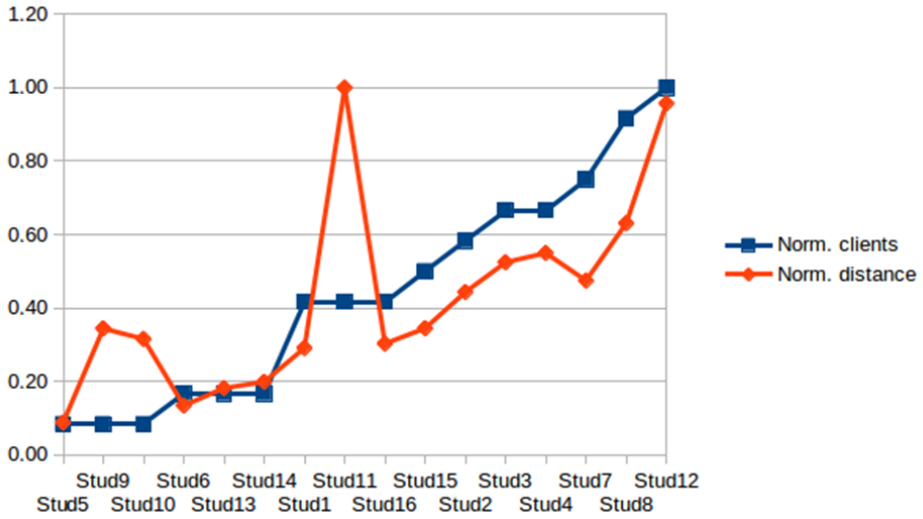

Figure 8. General group behaviour.

- *Validated*: from the teacher's point of view, the indicator for the number of clients attended complements the previous indicator (distance covered) when assessing students' interaction in the target language, since both indicators highlight the effort students had to make in order to accomplish the given task. Moreover, the data from Figure 8 show the more clients a student attended, the greater the amount of effort he/she was required to deliver (i.e., distance covered). In this context, it stands out that Stud6, Stud13 and Stud14 respectively covered 0.13, 0.18 and 0.20 metres (out of a normalised maximum of It 1.0) in order to attend the same normalised number of clients (0.17). The same applies to Stud3 and Stud4, who both attended the same normalised number of clients (0.67), covering 0.52 and 0.54 metres respectively (out of a normalised maximum of It 1.0). Nonetheless, there are several exceptions: the most outstanding case is Stud11, who covered the longest distance (1.0) while attending a below-average normalised number of clients (0.41). A similar difference can be observed in the case of Stud9 and Stud10. Therefore, the question of what might have caused the difference between the distance covered by each student and the number of orders (clients) attended is raised. A look at

the previous indicators explains the differences: on the one hand, both Stud9 and Stud10 only attended 1 client (see Table 2), after having played the game for 3-min (see Table 3), so their interaction was not long enough to be taken into account. On the other hand, Stud11 showed a significant interaction with the VW, while his/her performance was average: he/she attended 5 orders in the same time his/her game-partner (Stud12) was able to attend 12 orders, so he/she probably needed to approach his/her clients several times to make sure that his/her orders were delivered correctly

6. Discussions

From the language teachers' point of view, the implications of being able to assess students by using activity records extracted from VWs and an EUD programming language that allows them to customise their assessments are the following.

1. It allows language teachers to sustainably assess VW-based learning experiences by focusing not only on the final results according to the scores achieved by each learner, but also on how each individual learner interacted with the game. By extracting the records of students' activity while playing the game, it is possible to take into account how they performed with regard to other indicators. For example, if learners have to complete a task (deliver orders) in a given time and the teacher focuses only on the scores obtained, learners who are less skilled at playing video games [15] might score even lower by having a 100% success compared to other learners who have a 50% success and have moved much faster through the game attending more clients and delivering more orders.
2. It allows language teachers to think about designing learning experiences based on VWs by focusing on all the actions the learner has to take [12]. For example, it has been found that the fact that one learner had to run a longer distance compared to another learner in order to attend his/her clients was a disadvantage. Being able to extract this information allows the teacher firstly, to consider this feature for making a fairer assessment and secondly, to consider the mentioned aspect when redesigning the game to avoid that running a longer distance to complete a given task implies for the respective learner a disadvantage when competing with other learners who are not required to run the same distance.
3. Thanks to the use of an EUD programming language, the teacher can develop his or her own evaluation criteria rather than depending on VWs and video games which, although they could provide some indicators based on students' activity records, are usually tailor-made in line with the interests of those (teacher or computer specialist) who have created the game [22]. This can be illustrated by an example from the VW-based game (GEFE) that has been implemented for the current study. For example, if the game had only provided the score of each student and the time spent to perform the given learning task, it would not have been possible to refine the indicators by taking into account aspects such as the number of orders delivered correctly or the distance run, which have later been proven to be relevant for the current study. Finally, it is important to note that the indicators used in this study to assess the students' language skills (i.e., vocabulary knowledge and interaction in the target language) have been endorsed by the experience of the language teacher, who participated in the current study and who has refined them on the basis of her personal experience as well as observation of students' game performance. Taking into consideration the results from the current study the authors consider that the added value of using tools as the one discussed in this paper lies in the fact that it will be the language teacher himself/herself who will be able to use, adapt, discard or redesign his/her indicators through the use of the EUD programming language.

Finally, it is important to note that the indicators used in this study to assess the language skills of students, i.e., vocabulary knowledge and interaction in the target language, have been endorsed by the experience of the language teacher, who refined them on the basis of observation of the activity and her experience. The power of the use of

these tools lies in the fact that it will be the language teacher himself/herself who will be able to use, adapt, discard or redesign his/her indicators through the use of the EUD programming language.

7. Conclusions

In terms of in-game assessment, still very little research has been done to help language teachers analysing and assessing students' performance when learning through VW-based environments. While external game features such as game scores and play duration have turned out to be unfair in some evaluations (especially when randomisation comes into play), other more internal aspects such as students' interaction and the effort made to complete the given game task successfully, could reveal some interesting information on students' learning process and the language knowledge they have acquired.

The current study aims to help language teachers assessing students' performance when learning through VW-based games, by providing them with some valuable tools to sustainably analyse and assess students' language skills. With this purpose in mind, the paper suggests the use of both, an evidence-based method as well as an EUD programming language. To corroborate the validity of the proposed method for students' language assessment, a case study with students from an undergraduate German foreign language course was conducted.

The findings of this study reveal that the use of VW interaction logs can help teachers to sustainably conduct a fine-grained analysis of students' performance and facilitate their assessment in VW-based game environments:

1. Firstly, a specific EUD programming language allows teachers to easily retrieve objective indicators from students' logs together with the use of a range of visualisation methods (i.e., VW tracking map and general data report). This way, they can sustainably (that is, independently from the number of students) gather valuable information on students' task performance and the skills involved.
2. Secondly, an evidenced-based method allows teachers to refine their initial assessment criteria and thus implement a more comprehensive assessment method. This refinement is especially helpful when considering differences in terms of students' performance due to external factors such as the randomisation of game tasks. The analysis of students' results illustrated that such factors could significantly affect their game performance, since the effort required to fulfil the given task could differ in each case and game session.

Future work needs to focus on more specific aspects such as students' use of the foreign language in terms of specific structures, communication strategies etc. to be included in the assessment process. Apart from refining language assessment, the authors intend to extend their study to other professional fields and skills e.g., generic skills such as teamwork or leadership skills, in order to provide teachers from a wide range of areas with the means to easily assess their students' performance when using games for learning.

Author Contributions: Conceptualisation, J.M.D., M.P.-D. and A.B. (Anke Berns); methodology, A.B. (Antonio Balderas) and M.P.-D.; software, A.B. (Antonio Balderas); validation, J.M.D., and D.C.; formal analysis, A.B. (Anke Berns), and D.C.; investigation, M.P.-D. and A.B. (Antonio Balderas); writing—original draft preparation, M.P.-D., A.B. (Anke Berns) and A.B. (Antonio Balderas); writing—review and editing, A.B. (Antonio Balderas). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Spanish National Research Agency (AEI), through the projects VISAIGLE (TIN2017-85797-R) and DeepBio (TIN2017-85727-C4-3-P) with ERDF funds.

Institutional Review Board Statement: Ethical review and approval were waived for this study, due to not involving personally identifiable nor sensitive data.

Informed Consent Statement: Student consent was waived due to not involving personally identifiable nor sensitive data.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://doi.org/10.6084/m9.figshare.13491240.

Acknowledgments: We would further like to thank Raúl Gómez Sánchez, Francisco Rodríguez and Owayss Kabtoul for their technical support, as well as the Oficina de Software Libre y Conocimiento Abierto (OSLUCA) at the University of Cadiz for their much-valued support during the entire project.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CEFR	Common European Framework of Reference
EUD	End-User Development
ECTS	European Credit Transfer and Accumulation System
GEFE	German Expert and Fast waitEr
IT	Information Technology
DSL	Domain-Specific Language
MMORPG	Massively Multiplayer Online Role-Playing Games
MUVE	Multiuser Virtual Environment
VW	Virtual World
VWQL	Virtual World Query Language

References

1. Molka-Danielsen, J.; Deutschmann, M. *Learning and Teaching in the Virtual World of Second Life*; Tapir Academic Press: Trondheim, Norway, 2009.
2. Stevens, V. Second Life in education and language learning. *TESL-EJ* **2006**, *10*, 1–4.
3. Svensson, P. Virtual worlds as arenas for language learning. In *Language Learning Online: Towards Best Practice*; CRC Press: Boca Raton, FL, USA, 2003; Chapter 7; pp. 123–142.
4. Garrido-Iñigo, P. Viabilidad de las plataformas virtuales en la enseñanza de una lengua extranjera. *Rev. Investig. Educ.* **2012**, *10*, 129–140.
5. Berns, A.; Palomo-Duarte, M.; Dodero, J.M.; Valero-Franco, C. Using a 3D online game to assess students' foreign language acquisition and communicative competence. In Proceedings of the European Conference on Technology Enhanced Learning, Paphos, Cyprus, 17–21 September 2013; Springer: Berlin/Heidelberg, Heidelberg, 2013; pp. 19–31.
6. Garrido-Iñigo, P.; Rodríguez-Moreno, F. The reality of virtual worlds: Pros and cons of their application to foreign language teaching. *Interact. Learn. Environ.* **2015**, *23*, 453–470. [CrossRef]
7. Caballero-Hernández, J.A.; Palomo-Duarte, M.; Dodero, J.M. Skill assessment in learning experiences based on serious games: A systematic mapping study. *Comput. Educ.* **2017**, *113*, 42–60. [CrossRef]
8. Berns, A.; Gonzalez-Pardo, A.; Camacho, D. Game-like language learning in 3-D virtual environments. *Comput. Educ.* **2013**, *60*, 210–220. [CrossRef]
9. Chotipaktanasook, N.; Reinders, H. A massively multiplayer online role-playing game and its effects on interaction in the second language: Play, interact, and learn. In *Handbook of Research on Integrating Technology into Contemporary Language Learning and Teaching*; IGI Global: Hershey, PA, USA, 2018; pp. 367–389.
10. Griffiths, M.D. The educational benefits of videogames. *Educ. Health* **2002**, *20*, 47–51.
11. Reinders, H.; Wattana, S. Affect and willingness to communicate in digital game-based learning. *ReCALL J. Eurocall* **2015**, *27*, 38. [CrossRef]
12. Bellotti, F.; Kapralos, B.; Lee, K.; Moreno-Ger, P.; Berta, R. Assessment in and of serious games: An overview. *Adv. Hum. Comput. Interact.* **2013**, *2013*, 136864. [CrossRef]
13. Rama, P.S.; Black, R.W.; Van Es, E.; Warschauer, M. Affordances for second language learning in World of Warcraft. *ReCALL J. Eurocall* **2012**, *24*, 322. [CrossRef]
14. Wang, C.P.; Lan, Y.J.; Tseng, W.T.; Lin, Y.T.R.; Gupta, K.C.L. On the effects of 3D virtual worlds in language learning—A meta-analysis. *Comput. Assist. Lang. Learn.* **2020**, *33*. [CrossRef]
15. Peterson, M. Digital gaming and second language development: Japanese learners interactions in an MMORPG. *Digit. Cult. Educ.* **2011**, *3*, 56–73.
16. Peterson, M. EFL learner collaborative interaction in Second Life. *ReCALL* **2012**, *24*, 20–39. [CrossRef]
17. Balderas, A.; Berns, A.; Palomo-Duarte, M.; Dodero, J.M.; Gómez-Sánchez, R.; Ruiz-Rube, I. A domain specific language to retrieve objective indicators for foreign language learning in virtual worlds. In Proceedings of the 3rd International Conference on Technological Ecosystems for Enhancing Multiculturality, Porto, Portugal, 7–9 October 2015; pp. 675–680.

18. Sylvén, L.K.; Sundqvist, P. Gaming as extramural English L2 learning and L2 proficiency among young learners. *ReCALL* **2012**, *24*, 302–321. [CrossRef]
19. Minovic, M.; Štavljanin, V.; Milovanovic, M. Educational games and IT professionals: Perspectives from the field. *Int. J. Hum. Cap. Inf. Technol. Prof. (IJHCITP)* **2012**, *3*, 25–38. [CrossRef]
20. Balderas, A.; Berns, A.; Palomo-Duarte, M.; Dodero, J.M.; Ruiz-Rube, I. Retrieving Objective Indicators from Student Logs in Virtual Worlds. *J. Inf. Technol. Res. (JITR)* **2017**, *10*, 69–83. [CrossRef]
21. Palomo-Duarte, M.; Berns, A.; Yañez Escolano, A.; Dodero, J.M. Clustering analysis of game-based learning: Worth it for all students? *J. Gaming Virtual Worlds* **2019**, *11*, 45–66. [CrossRef]
22. Hsiao, I.Y.; Lan, Y.J.; Kao, C.L.; Li, P. Visualization analytics for second language vocabulary learning in virtual worlds. *J. Educ. Technol. Soc.* **2017**, *20*, 161–175.
23. Cruz-Benito, J.; Therón, R.; García-Peñalvo, F.J.; Lucas, E.P. Discovering usage behaviors and engagement in an Educational Virtual World. *Comput. Hum. Behav.* **2015**, *47*, 18–25. [CrossRef]
24. Minović, M.; Milovanović, M.; Šošević, U.; González, M.Á.C. Visualisation of student learning model in serious games. *Comput. Hum. Behav.* **2015**, *47*, 98–107. [CrossRef]
25. Gómez-Aguilar, D.A.; Hernández-García, Á.; García-Peñalvo, F.J.; Therón, R. Tap into visual analysis of customization of grouping of activities in eLearning. *Comput. Hum. Behav.* **2015**, *47*, 60–67. [CrossRef]
26. Gonzalez-Pardo, A.; Rosa, A.; Camacho, D. Behaviour-based identification of student communities in virtual worlds. *Comput. Sci. Inf. Syst.* **2014**, *11*, 195–213. [CrossRef]
27. Alvarado, S.H.; Cortiñas, A.; Luaces, M.R.; Pedreira, O.; Places, Á.S. Developing Web-based Geographic Information Systems with a DSL: Proposal and Case Study. *J. Web Eng.* **2020**, *19*, 167–194. [CrossRef]
28. Oates, B.J. *Researching Information Systems and Computing*; Sage: Thousand Oaks, CA, USA, 2005.
29. Balderas, A.; Palomo-Duarte, M.; Dodero, J.M.; Ibarra-Sáiz, M.S.; Rodríguez-Gómez, G. Scalable authentic assessment of collaborative work assignments in wikis. *Int. J. Educ. Technol. High. Educ.* **2018**, *15*, 40. [CrossRef]
30. Balderas, A.; De-La-Fuente-Valentin, L.; Ortega-Gomez, M.; Dodero, J.M.; Burgos, D. Learning management systems activity records for students' assessment of generic skills. *IEEE Access* **2018**, *6*, 15958–15968. [CrossRef]
31. Balderas, A.; Caballero-Hernández, J.A.; Dodero, J.M.; Palomo-Duarte, M.; Ruiz-Rube, I. Model-Driven Skills Assessment in Knowledge Management Systems. *J. Web Eng.* **2019**, *18*, 353–380. [CrossRef]
32. Gómez Sánchez, R. Evalsim: Sistema para la Extracción de Indicadores de Interacción en Mundos Virtuales. (Degree Final Project, Universidad de Cádiz). 2017. Available online: http://hdl.handle.net/10498/24104 (accessed on 22 December 2020). (In Spanish)