School of Information Systems

Taxonomy of Human Actions for Action-based Learning Assessment in Virtual Training Environments

Ali Fardinpour

This thesis is presented for the Degree of
Doctor of Philosophy
Of
Curtin University

December 2016

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made. This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Human Ethics (For projects involving human participants/tissue, etc) The research presented and reported in this thesis was conducted in accordance with the National Health and Medical Research Council National Statement on Ethical Conduct in Human Research (2007) – updated March 2014. The proposed research study received human research ethics approval from the Curtin University Human Research Ethics Committee (EC00262), Approval Number # IS_14_24, IS_14_15, and IS_14_04.

Signature:………………………………………….

Date:………………………...

Acknowledgments

I wish to take this opportunity to express my sincere gratitude to my supervisor, Dr. Torsten Reiners, who always had faith in me and always treated me like a colleague and a researcher. He is a tough critic, my fantastic friend, and my motivation in challenging times. I also thank my co-supervisors, A/Professor David Gibson and Dr. Sue Gregory who supported me with their intellectual inspiration and guidance.

I acknowledge and respect my mentor and first supervisor, and co-supervisor after his retirement, Professor Heinz Dreher, who was always there when I needed him the most, supporting me, guiding me and nurturing the researcher and scholar in me. He stood beside me and supported me in every way possible and encouraged me to fight for what I deserved. I also greatly appreciate the friendship, kindness and support extended to me by his fantastic daughter, Naomi.

I acknowledge the constant support of the staff members at the School of Information Systems and Higher Degree by Research office who were always kind and supportive.

Dedication

To my parents who simply gave me all they had, all the love in the world, all the trust one can have in a person. And all they have asked of me was that I pass that to others.

Abstract

This research has explored how the actions performed by learners can be formally represented to create consistency when comparing these actions with expert reference solutions, in order to generate an automated post-performance formative feedback. The assessment of learners' performed actions was intended to overcome the inability of general assessment methods to represent the acquired applicable knowledge or skills. The required consistency in the assessment process necessitated a formal representation of the actions, and the lack of such representation needed to be addressed.

This research developed a taxonomy of human actions for use in an assessment methodology based on performed actions. Action-based Learning Assessment Methodology provides a framework for the comparison of learners' actions to reference solutions and the generation of an automated post-performance formative feedback. A crucial requirement of the assessment methodology is the delivery of a consistent result. Therefore, the BEHAVE taxonomy provides a standard classification of human actions and a formal syntax to represent the actions as computer-readable codes similar to the performed action in the real-life or simulated setting.

The Design Science Research approach was used as the research methodology for this thesis. The assessment methodology and taxonomy were developed by combining primary and secondary research, collecting data via an expert opinion survey, and a literature review. Also, the expertise of the researcher played a substantial role in the development process. As part of the Design Science Research process, BEHAVE was evaluated for internal and external validity using the survey of experts, card sorting test, performance coding experiment, and participant feedback.

The results of the evaluations showed that BEHAVE is a valid taxonomy of human actions, both internally and externally. The survey of expert opinion demonstrated the face validity of BEHAVE; the exclusiveness and exhaustiveness of the taxonomy classes were shown by cluster analysis of the card sorting test, and the performance coding experiment and participant feedback provided the external validation of BEHAVE.

Publications Associated with This Thesis

Publications: International Conferences

Fardinpour, A., Dreher, H. (2012). Towards developing automated assessment in virtual learning environments, *The IADIS International Conference on Internet Technologies & Society*, 211-216.

Fardinpour, A., Reiners, T., Dreher, H. (2013). Action-based Learning Assessment Method (ALAM) in Virtual Training Environments. In M. Gosper, J. Hedberg, H. Carter (Eds.) *Electric Dreams. Proceedings Ascilite Sydney 2013* (pp. 267-276). Sydney, NSW: Macquarie University.

Fardinpour, A., Reiners, T. (2014). The Taxonomy of Goal-oriented Actions in Virtual Training Environments, *Procedia Technology*, 13, 38 - 46.

Publications: Book Chapter

Fardinpour, A., Reiners, T., Wood, L. C. (in press). Action-based Learning Assessment Method in Virtual Training Environments. In: Gregory, S., Wood, D., Scutter, S., Parsons, D., Atkins, C., Jacka, L. and Neuendorf, P. (eds.) *"Virtual Worlds: Facilitating Student Engagement, Creativity and Intercultural Awareness through Authentic Learning Experiences"*.

Table of Contents

Acknowledgments III

Dedication IV

Abstract V

Publications Associated with This Thesis VI

Publications: International Conferences VI
Publications: Book Chapter VI

Table of Contents VII

List of Figures XI

List of Tables XIII

List of Abbreviations XVI

Chapter 1: Introduction 1

1.1. Introduction 1

1.2. Research Scope 3

1.3. Research Motivation 4

1.4. Research Purpose 6

1.5. Research Significance 7

1.5.1. Contribution to Theory 7
1.5.2. Contribution to Practice 7

1.6. Organisation of the Thesis 8

Chapter 2: Literature Review 9

2.1. Introduction 9

2.2. Action-based Learning 12

2.2.1. Experiential Learning 14
2.2.2. Problem-based Learning 16
2.2.3. Authentic Learning 17
2.2.4. Inquiry-based Learning 19
2.2.5. Scenario-based Learning 19
2.2.6. Adventure Learning 20

2.3. Learning Assessment 21

2.3.1. Summative and Formative Assessment 21
2.3.2. Feedback 25
2.3.3. Authentic Assessment 27
2.3.4. Automated Assessment 30

2.4. Virtual Environments for Learning and Assessment 32

2.5. Taxonomy of Human Actions 38

2.5.1. Human Action, Activity, Behaviour, Task, and Performance 39
2.5.2. Context and Interpretation 41
2.5.3. Taxonomy and Different Fields of Use 42
2.5.4. Review of Current Taxonomies 43
2.5.4.1. Actions 43
2.5.4.2. Activities 44
2.5.4.3. Gestures 46
2.5.4.4. Virtual and Computerised Environments 47
2.5.5. Task, Skill, Behaviour, and Performance 50
2.5.6. Discussion 52

2.5.7. Evaluation of a Taxonomy 57

2.6. Summary and Implications 59

Chapter 3: Research Methodology 62

3.1. Introduction 62

3.2. Research Question 62

3.3. Research Aim and Objectives 63

3.3.1. Research Aim 63

3.3.2. Research Objectives 63

3.4. Introduction to Research Design and Methodology 65

3.4.1. Research Paradigms 65

3.4.1.1. Postpositivist 66

3.4.1.2. Constructivist 66

3.4.1.3. Transformative 67

3.4.1.4. Pragmatic 67

3.4.1.5. Applied Research Paradigm 67

3.4.2. Research Methodologies 70

3.4.3. DSR: Different Frameworks 73

3.5. Applied Research Methodology 81

3.5.1. Step 1: Problem Identification and Motivation 82

3.5.2. Step 2: Define the Objectives and Anticipated Significance of Solution 83

3.5.3. Step 3: Design and Development 83

3.5.4. Step 4: Demonstration and Evaluation 85

3.5.5. Step 5: Communication 85

3.6. Evaluation and Data Analysis Methods 85

3.6.1. Survey for Expert Opinion: Virtual Worlds and Industry Experts 89

3.6.2. Card Sorting Test 89

3.6.3. Performance Coding Experiment and Participant Feedback 91

3.7. Participants 93

3.8. Research Ethics 95

3.9. Summary 95

Chapter 4: Action-based Learning Assessment Methodology 97

4.1. Introduction 97

4.2. Action-based Learning Assessment Methodology 98

4.2.1. ALAM Supports Authentic Assessment 100

4.2.2. ALAM Development Process 103

4.3. The ALAM Main Characteristics and Components 104

4.3.1. Mapping the Actions through a Formal Syntaxes 104

4.3.2. Comparison to Multiple Expert Reference Solutions 105

4.3.3. Feedback for Learning 106

4.4. Assessment System Conceptual Model 108

4.5. Summary 111

Chapter 5: BEHAVE: Taxonomy of Human Actions 112

5.1. Introduction 112

5.2. BEHAVE taxonomy 113

5.2.1. Definitions 114

5.2.2. BEHAVE Taxonomy: Action Attributes and Syntax 117

5.3. Example Scenarios 119

5.3.1. Applying the Behave Taxonomy to HTA Example 119
5.3.2. Applying the BEHAVE Taxonomy: Other Example Scenarios 123
5.3.2.1. Cooking Dinner for the Family (Squash Soup for 4) 123
5.3.2.2. Supporting Copper Rebar in the Lathe Chuck 126

5.4. Actions in Different Contexts 128

5.5. (English) Language Effects on the Taxonomy of Human Actions 129

5.6. Need and Use of the BEHAVE Taxonomy 131

5.7. BEHAVE Application in ALAM 132
5.7.1. Performance Codifier Engine 132
5.7.2. Comparison Engine 136
5.7.3. Feedback Compiler Engine 137

5.8. Summary 137

Chapter 6: Data and Analysis 139

6.1. Introduction 139

6.2. Survey Results 139
6.2.1. Survey Results for All the Respondents 140
6.2.1.1. Importance of Each Level of the Taxonomy 140
6.2.1.1.1. Friedman's test for prioritising the taxonomy's levels 143
6.2.1.2. The Importance of Functional Classes, BEHAVE, Multiple Expert Solutions, And Formative Feedback 144
6.2.1.3. Known Current Taxonomies of Human Actions 146
6.2.2. The Results of VW and Industry Experts 147
6.2.2.1. Virtual World Experts: Questions 1-3 147
6.2.2.1.1. Chi-square goodness of fit test for questions 1-3: 147
6.2.2.1.2. Friedman's test for prioritising the taxonomy levels 151
6.2.2.2. Industry Experts: Questions 1-3 152
6.2.2.2.1. Chi-square goodness of fit test for questions 1-3: 152
6.2.2.2.2. Friedman's test for prioritising the taxonomy levels 155
6.2.2.3. Both Groups' Results for the Questions 4-16 156
6.2.2.3.1. Chi-Square Test of Homogeneity 159

6.3. Card Sorting Test Results 161
6.3.1. Cluster Analysis 161
6.3.2. Similarity of Calculated and Perceived Clusters 166
6.3.3. Partitioning Validity: Dispersion of Data 166
6.3.4. Pearson Correlation Test 167

6.4. Performance Coding Experiment Results 168
6.4.1. Normality Tests 169
6.4.2. Coded Actions Analysis 173
6.4.3. Participants' Feedback Analysis 175

Chapter 7: Discussions 179

7.1. Introduction 179

7.2. Survey for Expert Opinion: VW and Industry Experts 179

7.3. Card Sorting Test 181

7.4. Performance Coding Experiment 182
7.4.1. Participants' Coding 182
7.4.2. Participants' Feedback 184

7.5. Taxonomy Evaluation 184

7.6. Summary 187

Chapter 8: The Past, Present and Future 188

8.1. Introduction 188

8.2. Past 188

8.3. Present 189

8.3.1. Research Question and Objectives 189

8.3.2. Research Practicality 191

8.3.3. What Are The Constraints? 192

8.4. Future 192

8.5. Conclusion 193

REFERENCES 195

Appendix 1: Survey questionnaire 215

Appendix 2: Willis input-output three level model 218

Appendix 3: Berliner classificatory scheme 219

Appendix 4: Meister taxonomy 220

Appendix 5: Bennett's Semantic Classificatory Approach (1971) 221

Appendix 6: Task representing different ability categories 222

Appendix 7: Task and skill classification system 227

Appendix 8: RATaC taxonomy 229

Appendix 9: HAR process as applied to VE system development 230

Appendix 10: Performance Shaping Factor taxonomies 231

Appendix 11: RATaC task taxonomy 233

Appendix 12: Maintenance task taxonomy 234

Appendix 13: Karam & Schraefel categories of gestures 235

Appendix 14: Data analysis tables and figures 237

Appendix 15: Copyright Documents 252

List of Figures

Figure 2.1: Change of learning environment, communication, and teaching ------------------10
Figure 2.2: A volunteer from C3KC teaches Sarah to solder -----------------------------------11
Figure 2.3: Figure 2.3: Foundational scholars of experiential learning theory (Kolb & Kolb, 2012) ---14
Figure 2.4: Figure 2.4: PBL Process by Hung et al. (2008) -------------------------------------17
Figure 2.5: Figure 2.5: Spacewalk in the Integrated EVA/RMS Virtual Reality Simulator Facility at Johnson Space Centre ---33
Figure 2.6: Figure 2.6: Agent A flipping the switch (Tuomela, 1977) ------------------------41
Figure 2.7: Hierarchical levels of activity (Kuutti, 1995) ---------------------------------------45
Figure 2.8: Manually-designed functional ADL taxonomy (Pirsiavash & Ramanan, 2012) 46
Figure 3.1: Initial and new aim and objectives ---64
Figure 3.2: Design Science Research process (DSRP) model ----------------------------------74
Figure 3.3: DSR process (Offermann et al., 2009) ---74
Figure 3.4: Research DSR framework developed for this research ----------------------------82
Figure 3.5: DRS process used in this research ---84
Figure 3.6: Choosing appropriate statistic tests (Kothari, 2004) -------------------------------88
Figure 3.7: Decision-making chart for choosing the best mean comparison test ------------93
Figure 4.1: ALAM development process ---104
Figure 4.2: A mock-up of ALAM formative feedback based on example in Section 5.3.2.2 ---107
Figure 4.3: Action-based Learning Assessment System Conceptual Model -----------------109
Figure 4.4: Action-based Learning Assessment Systems input interface wireframe --------109
Figure 4.5: Action-based Learning Assessment Systems database interface wireframe ---110
Figure 4.6: Action-based Learning Assessment Systems output interface wireframe ------111
Figure 5.1: The Hierarchical Levels and Classes of BEHAVE -------------------------------114
Figure 5.2: HTA of the task 'boil kettle' (Created by TaskArchitect3 software) ------------120
Figure 6.1: Respondents' views about the importance of the Goal Act ----------------------142
Figure 6.2: Respondents' views about the importance of Constitutive Acts ----------------143
Figure 6.3: Respondents' views about the importance of Functional Act -------------------143
Figure 6.4: Participants' responses to questions 4 ---238
Figure 6.5: Participants' responses to questions 5 ---238
Figure 6.6: Participants' responses to questions 6 ---238
Figure 6.7: Participants' responses to questions 7 ---238
Figure 6.8: Participants' responses to questions 8 ---238
Figure 6.9: Participants' responses to questions 9 ---238

Figure 6.10: Participants’ responses to questions 10 ---238
Figure 6.11: Participants’ responses to questions 11 ---238
Figure 6.12: Participants’ responses to questions 12 ---239
Figure 6.13: Participants’ responses to questions 13 ---239
Figure 6.14: Participants’ responses to questions 14 ---239
Figure 6.15: Distribution of respondents' views about their level of familiarity with the use of taxonomy --146
Figure 6.16: Respondents’ views about the importance of the Goal Act ---------------------149
Figure 6.17. Respondents’ views about the importance of Constitutive Acts ----------------150
Figure 6.18. Respondents’ views about the importance of Functional Acts -----------------151
Figure 6.19: Respondents’ views about the importance of the Goal Act ---------------------154
Figure 6.20. Respondents’ views about the importance of Constitutive Acts ----------------154
Figure 6.21. Respondents’ views about the importance of Functional Acts -----------------155
Figure 6.22: Comparing both groups of respondents’ views on questions 4------------------243
Figure 6.23: Comparing both groups of respondents’ views on questions 5 -----------------243
Figure 6.24: Comparing both groups of respondents’ views on questions 6 -----------------243
Figure 6.25: Comparing both groups of respondents’ views on questions 7 -----------------243
Figure 6.26: Comparing both groups of respondents’ views on questions 8 -----------------243
Figure 6.27: Comparing both groups of respondents’ views on questions 9 -----------------243
Figure 6.28: Comparing both groups of respondents’ views on questions 10 ---------------244
Figure 6.29: Comparing both groups of respondents’ views on questions 11 ---------------244
Figure 6.30: Comparing both groups of respondents’ views on questions 12 ---------------244
Figure 6.31: Comparing both groups of respondents’ views on questions 13 ---------------244
Figure 6.32: Comparing both groups of respondents’ views on questions 14 ---------------244
Figure 6.33: Comparing the distribution of respondents' level of familiarity with the use of human actions taxonomy --158
Figure 6.34: Scatterplots’ matrix for card sorting test --162
Figure 6.35: K-mean cluster analysis for five and seven clusters -----------------------------167
Figure 6.36: Sorted items colour map --251
Figure 6.37: Boxplot of number of actions in video group ------------------------------------170
Figure 6.38: Boxplot of number of actions in written narration group -----------------------171
Figure 6.39: The Videotaped Scenario Group Results ---174
Figure 6.40: The Written Scenario Group Results --175
Figure 6.41: Relative frequencies for descriptive words used by participants ---------------178

List of Tables

Table 2.1: Feedback Types Arrayed Loosely by Complexity (Shute, 2007) ------------------25

Table 2.2: A comparison of authentic and traditional assessment (Herrington & Herrington, 1998; Wiggins, 1990) --28

Table 2.3: Early developed VTEs --34

Table 2.4: Four Gesture Classification Schemes (McNeill, 1992) ------------------------------47

Table 2.5: Definition of Avatar Capabilities Model (by permission from Chodos et al., 2014) --50

Table 2.6: Published taxonomies of human actions, performances, tasks, and behaviour ---54

Table 2.7: Taxonomy validating methods in different research --------------------------------59

Table 3.1: The four belief systems addressed by Mertens (2015: p.11) for the four common paradigms --69

Table 3.2: The four belief systems for the applied paradigms in this research ----------------70

Table 3.3: DSR steps in literature (steps are in original titles used by each author) ----------77

Table 3.4: Variables and values for the evaluation of DSR artefacts (Cleven et al., 2009) --79

Table 3.5: Mapping artefacts and evaluate methods from the literature (Peffers et al., 2012) - --80

Table 3.6: Evaluation methods and data analysis tools --93

Table 5.1: 'boil kettle' example code by BEHAVE syntax ------------------------------------122

Table 5.2: 'Preparing for cooking' example code by BEHAVE syntax ---------------------125

Table 5.3: 'Put the rebar in the chuck and fix the tailstock' example code by BEHAVE syntax ---127

Table 5.4: Different meanings of an action leading to different classifications – results from the card sorting test explained in Section 6.4 ---129

Table 5.5: 'Boil Kettle' ALAM assessment scenario code by BEHAVE syntax ------------134

Table 6.1: Chi-square test descriptive statistics for the Goal Act -----------------------------141

Table 6.2: Chi-square test descriptive statistics for the Constitutive Acts --------------------141

Table 6.3: Chi-square test descriptive statistics for the Functional Acts ----------------------141

Table 6.4: Chi-square test statistics and results for the Goal Act -----------------------------141

Table 6.5: Chi-square test statistics and results for the Constitutive Acts --------------------142

Table 6.6: Chi-square test statistics and results for the Functional Acts ----------------------142

Table 6.7: Friedman's test's descriptive statistics --144

Table 6.8: Friedman's test's results --145

Table 6.9: Descriptive statistics ---237

Table 6.10: Chi-square statistics and results for questions 4 to 14 ----------------------------145

Table 6.11: Binomial test results for all respondents --146

Table 6.12: Survey questions’ variance ---240

Table 6.13: Chi-square test descriptive statistics for the Goal Act -----------------------------148

Table 6.14: Chi-square test descriptive statistics for the Constitutive Acts ------------------148

Table 6.15: Chi-square test descriptive statistics for the Functional Acts --------------------148

Table 6.16: Chi-square test statistics and results --148

Table 6.17: Chi-square test statistics and results --149

Table 6.18: Chi-square test statistics and results --150

Table 6.19: Friedman’s test’s descriptive statistics ---151

Table 6.20: Friedman test’s results --151

Table 6.21: Chi-square test descriptive statistics for the Goal Act -----------------------------152

Table 6.22: Chi-square test descriptive statistics for the Constitutive Acts ------------------152

Table 6.23: Survey questions’ variance ---240

Table 6.24: Chi-square test descriptive statistics for the Functional Acts --------------------152

Table 6.25: Chi-square test statistics and results --153

Table 6.26: Chi-square test statistics and results --153

Table 6.27: Chi-square test statistics and results --153

Table 6.28: Friedman’s test’s descriptive statistics ---155

Table 6.29: Friedman test’s results --155

Table 6.30: Descriptive statistics of industry experts ---241

Table 6.31: Descriptive statistics of virtual worlds’ experts ------------------------------------242

Table 6.32: Chi-square statistics and results of industry experts -------------------------------156

Table 6.33: Chi-square statistics and results of VW experts -----------------------------------157

Table 6.34: Binomial test results comparison between both groups --------------------------159

Table 6.35: Chi-Square Test of Homogeneity results ---160

Table 6.36: Initial cluster centres for card sorting test ---163

Table 6.37: Iteration history and changes in cluster centres ------------------------------------245

Table 6.38: Final cluster centres that result after the stop command --------------------------164

Table 6.39: Distances between Final Cluster Centres ---164

Table 6.40: Number of cases in each cluster ---164

Table 6.41: Cluster memberships and distances from cluster centres -------------------------246

Table 6.42: X_{ij} , $\overline{X}_J$ results ---247

Table 6.43: $\left(X_{ij}-\overline{X}_J\right)^2$ results ---248

Table 6.44: $\left(X_{ij}-\overline{X_{ij}}\right)^2$ results --249

Table 6.45: Percentage of each clustered item ---250

Table 6.46: Correlation matrix ---168

Table 6.47: Descriptive statistics of number of actions in video group -----------------------169

Table 6.48: Kolmogorov-Smirnov test results for video group --------------------------------170

Table 6.49: Mean comparison test results in video group --------------------------------------171

Table 6.50: Kolmogorov-Smirnov test results for written narration group -------------------172

Table 6.51: Descriptive statistics of number of actions in written narration group ---------172

Table 6.52: Mean comparison of test results in written narration group ----------------------172

Table 6.53: Videotaped scenario group data ---173

Table 6.54: Written scenario group data ---174

Table 6.55: Concept frequency ---177

Table 7.1: Correlations between the BEHAVE classes based on Pearson correlation test results interpreted by Evans & Over (2013) suggestions ---------------------------------------182

List of Abbreviations

AC: Abstract Conceptualization

ACM: Avatar Capabilities Model

ADL: Activities of Daily Living

AE: Active Experimentation

AI: Artificial Intelligence

ALAM: Action-based Learning Assessment Methodology

BEHAVE: Basic Exploratory Human Actions in Virtual Environments

CE: Concrete Experience

CSCW: Computer-supported Cooperative Work

DSR: Design Science Research

ELT: Experiential Learning Theory

HAR: Human Abilities Requirements

HCI: Human-Computer Interaction

HTA: Hierarchical Task Analysis

MARS: Manual for the Ability Requirement Scales

PSF: Performance Shaping Factor

RATaC: Rapid Assessment of Tasks and Context

TTRAM: Task and Training Requirements Analysis Methodology

VE: Virtual Environment

VR: Virtual Reality

VTE: Virtual Training Environment

VW: Virtual World

Chapter 1: Introduction

1.1. Introduction

When investigating automated computerised assessment systems, several questions arise: can we assess people's knowledge based on their actions instead of by their response to general assessment methods such as multiple choice questions, short answers, or essays that are incapable of representing learned applicable knowledge or skills? How can performed actions best be assessed? Since an assessment alone cannot improve learners’ applicable knowledge unless they learn from their mistakes, detailed formative feedback is needed to provide people with the opportunity to learn from their mistakes. To create automated formative feedback, the computer will need to recognise the performed actions, possibly via comparison with standard or reference actions. Therefore, we need a classification of human actions.

However, none of the current classifications of human actions is exhaustive enough (Section 2.5) for the required task, as all these taxonomies were developed for a very specific research need without the flexibility allowing them to be used by other disciplines. Consequently, a new taxonomy of human actions, namely the BEHAVE taxonomy (Chapter 5), is developed in this research.

The Basic Exploratory Human Actions in Virtual Environments (BEHAVE) taxonomy is developed by combining primary and secondary research, collecting data via an expert opinion survey, and a literature review. However, the expertise of the researcher plays a substantial role in the development of the taxonomy. The literature research contributed to the foundations of the BEHAVE taxonomy, based on the theory of human actions (Goldman, 1970), Taxonomy of Embodied Actions for cooperative design in a distributed company (Robertson, 1997, 2000), Avatar Capabilities Model (Chodos et al., 2014), and other taxonomies as shown in Table 2.6. Following an intensive study of these sources, the researcher applied his expertise in ontology development to establish the BEHAVE taxonomy’s levels, classes, and the action coding syntax. Direct observation of human actions in both real and simulated environments also made a major contribution to the creation of the BEHAVE taxonomy. The primary data collected by means of a survey contributed information for development and evaluation purposes. From this point forward, the BEHAVE taxonomy will be referred to as BEHAVE.

BEHAVE classifies human actions according to three main levels and six classes. The levels are based on Action-based Learning Assessment Methodology's goals (Section 4.2); these levels are: the Goal Act, Constitutional Acts, and Functional Acts. The Goal Act comprises one or more Constitutive Acts, which consist of one or more Functional Acts. Functional Acts, as the most basic action level, are classified according to six different classes: Gestural, Responsive, Decisional, Operative, Constructional, and Locomotive (Section 5.2.1).

As an artefact of Design Science Research (DSR) Methodology, BEHAVE was continuously evaluated during the iterative phases of the research and shown repeatedly to be an internally and externally validated framework, using experts' opinion, a card sorting test, empirical performance coding experimentation, and participant feedback.

BEHAVE enables coded actions and their attributes to be classified in order to generate automated formative feedback in a virtual training environment. This feedback is the result of a new assessment method founded in this research, called the Action-based Learning Assessment Methodology (ALAM), which uses actions performed in a digital environment to assess a learner's performance in a (2D/3D) Virtual Training Environment (VTE) in order to solve a given problem or achieve a particular goal.

This research started with an initial question: "How can learners' goal-oriented actions and action-sequences be represented, analysed, and automatically assessed in virtual world environments?" To address this question, the main objective of the research was to "develop an Action-based Learning Assessment System in virtual worlds". ALAM was developed as the theoretical framework for this assessment system.

To create consistency in ALAM so as to generate feedback, ALAM needs a standard representation of the performed actions. At this stage, the application of the ALAM theoretical framework to an operational automated assessment system relies on the realisation of standard representation of the performed actions. Therefore, this research focuses mainly on the development of BEHAVE to facilitate this realisation by a standard classification of human actions, an action-attributes set, and a standard syntax for action codification. The nexus between ALAM and BEHAVE is explained in Section 5.7. The research question, aim, and objectives are explained in Chapter 3.

The research method used in the development of ALAM was chosen using the researcher's expertise and a comprehensive literature review (Chapter 4). The following steps show the development process:

1. Review the literature on various theories and methods of Action-based Learning, formative feedback, and automated assessment;
2. Develop the ALAM framework;
3. Determine the main characteristics;
4. Develop the conceptual model of an assessment system based on ALAM.

1.2. Research Scope

The theoretical method of assessment called ALAM is based on experiential learning theory (Kolb, 1984; Kolb & Kolb, 2012) in a new context of performance in virtual worlds. As ALAM uses the performed actions of the learners and experts to generate an automated feedback, the input information of performed actions needs to follow a standard format that includes syntax and descriptive characteristics. For complete automation of the assessment process, a computerised performance is favourable, such as in 2D and 3D simulation-based environments. Simulation-based environments have the advantages of availability, cost efficiency, and safety measures. Highly immersive virtual 3D environments are used more frequently in training rather than 2D training environments due to their high degree of similarity to real-life scenarios. These contexts are favourable for eliciting and capturing both novice and expert performances in digital form.

For the sake of exhaustiveness and transferability, BEHAVE is designed to create a standard classification of human actions that is independent of the environment in which the actions occur. The framework includes a syntax to structure the presentation of the performed actions, and an action-attributes set to describe the performed actions similar to real-life actions. A virtual training environment, which is the focus of the ALAM, provides a high degree of similarity to real-life actions; moreover, it can provide a more precise description of the performed actions.

1.3. Research Motivation

Since the early days of using virtual reality technologies to create special effects in movies, the concept of virtuality has changed over the course of time (Bricken, 1990; Krueger, 1991; Steuer, 1992; Fuchs & Guitton, 2011). Virtual reality has evolved from implementations with goggles and peripherals to persistent interactive multi-user virtual worlds being used pervasively in different disciplines including education (Klastrup, 2003; Bell, 2008; Twining, 2009; Gregory et al., 2013). Virtual worlds have become more sophisticated and specialised over time. 3D VTE is a specialised Virtual Learning Environment used for training purposes (Sections 2.5.3 and 2.5.4). In education, the popularity of 3D simulations in training has increased due to a number of factors such as safety, availability, and affordability. Industries such as mining, gas and oil, and medicine started have benefitted from these employee training environments.

Traditionally, training and induction sessions are often based on PowerPoint presentations and talks which can be static and even tiring for trainers, and can lead to boredom for the learner and failure to take in the information being presented. To verify the successful completion of these training sessions, students often undertake summative assessment tasks such as multiple choice tests, short answers, closed-end questions, or signing the induction papers. As most of these assessment methods are based on memorisation, the memorised knowledge can be quickly forgotten or ignored over time. Furthermore, all these traditional assessment methods evaluate learners' knowledge at the 'remembering' level of Bloom's taxonomy of learning (Bloom et al., 1956), which does not show the learners' ability to apply their knowledge in real-life situations. According to Bloom et al. (1956), learners memorise information as the first step in the learning process, but this process can be developed further to result in knowledge creation, and finally self-evaluation by the learner.

Action-based Learning (also known as experiential learning or learning-by-doing) was developed with the purpose of enabling the learners to apply their knowledge in real-world, problem-solving scenarios (Naidu & Bedgood, 2012). The Action-based Learning method requires learners to perform certain actions, such as actively building, creating, or drawing something, in order to achieve a goal or solve a given problem. In some learning scenarios, learners might experience content by watching a video clip to apply the learned skill in a decision-making process that is later examined or reflected upon (Naidu & Bedgood, 2012). There are different types of Action-based Learning depending on the learning purpose, including Problem-based Learning (Barrows & Tamblyn, 1980), Inquiry or Goal-based Learning (Schwab, 1960; Herron, 1971; Edelson et al., 1999; Wilhelm & Wilhelm, 2010), Scenario-based Learning (Naidu, 2010), and Adventure Learning (Doering, 2006). Although the Action-based Learning method is frequently used in education, secondary assessment methods (e.g. tests, quizzes, opinion surveys) are still used to assess learners' knowledge, while direct assessment of learners' actions is often neglected. Assessment of performed actions can contribute to the learning process and enable learners to learn from their actions if used as an additional means of an overall assessment approach in a Learning-by-Doing method (Fardinpour et al., 2013). The lack of an action-focused assessment method in Action-based Learning, and the potential to use (2D/3D) VTE to facilitate the capture of data from digital actions, have motivated the development of ALAM, supported by a new taxonomy of human actions, BEHAVE. With ALAM, the performed actions of learners need to be mapped and processed so that an automated feedback can be generated. As the ALAM framework relies on the comparison of the learners' performance with multiple expert performances, learners' actions need to be mapped and processed using a standard representation method to facilitate the assessment process (Chapter 4).

Several disciplines have created taxonomies of human actions including task analysis (Swezey et al., 1998; Salmon et al., 2008), computer-supported cooperative work (Robertson, 1997, 2000), Avatar actions in VW (Hurst, 2011), Virtual interactions (Cappella & Pelachaud, 2002), and gestures (McNeill, 1992; Zhang et al., 2010). One important common factor across these disciplines is the need to order and arrange human actions and behaviours into groups or sets on the basis of their relationships and allocate any additional, previously unidentified actions to the correct class, once such classes have been established by prior classification (Simpson, 1961; Fleishman et al., 1984). Taxonomies and classifications facilitate the recognition and processing of the recorded information by different computerised systems for applications such as artificial intelligence. An investigation of the literature found that embodiment[1] and virtualisation are not as well-known as other areas (e.g. error recognition, hierarchy task analysis, and performance analysis). Computer-supported Cooperative Work (CSCW) and Action Learning are the only areas that use the classification of human actions in virtual environments with a business focus, but to date there is no exhaustive taxonomy in simulation-based fields (Section 2.4.4) (Goldman, 1970; Fleishman et al., 1984; Robertson, 1997, 2000; Cappella & Pelachaud, 2002; Stone, 2004; Pirsiavash & Ramanan, 2012). Furthermore, most of the developed taxonomies have a specific focus and cannot be extended and adapted. Therefore, there is the need for a taxonomy that is not constrained to an original application, but is open and flexible, allowing the extension, adaptation, and transfer to other scenarios and contexts.

This research project focuses on the development of such a taxonomy as an enabler of an automated assessment framework. One practical outcome of this focus is that it will contribute to the future development of assessment software, using the ALAM framework that can map digital actions in a virtual training environment, compare the actions to a set of reference solutions and generate automated formative feedback for learners to evaluate their performance so they can learn from their mistakes. The BEHAVE framework is a taxonomy of human actions that will support any field of human actions studies such as task analysis and video recognition, regardless of whether their environment is real or simulated.

1.4. Research Purpose

This study aims to develop a taxonomy to codify performed human actions to achieve an overall goal in an Action-based Learning scenario that requires solving a given problem. To achieve this aim, several objectives will be met: scrutinising different taxonomies and classifications of human actions in various disciplines; classifying human actions according to different levels and classes; designing a set of action attributes in order to describe an action; designing a precise syntax for codifying performed actions; and finally, applying taxonomically codified actions to generate assessment and feedback.

[1] "VEs engage the body as kinaesthetic input via the specialised interface devices that not only permit but require bodily actions to be performed sensorially, kinaesthetically, proprioceptively – within a full 3D spatial, yet virtual construct" (Morie, 2007, p. 126).

1.5. Research Significance

1.5.1. Contribution to Theory

This research contributes to the theory of learning and assessment by enabling the automated assessment of goal-oriented actions and learning at the highest levels of Bloom's taxonomy (Bloom et al., 1956) useful for the demonstration of knowledge. Automated feedback generation requires real-life actions to be reflected by their computerised equivalent. As the ALAM framework uses (2D/3D) VTE as the assessment environment, learners can repeatedly perform in a simulated version of a real-life learning scenario in order to improve their skills. The simulation-based environment allows a computer program to recognise the actions digitally, classify and process them, and create a formative feedback that helps learners to learn from their mistakes. To be able to classify the performed actions and describe them like their real-life action counterparts, BEHAVE uses different attributes, and structures the actions and their descriptive attributes with a formal syntax. The development of BEHAVE contributes to the theory of human actions of Alvin Goldman (1970) (Section 2.3.4) since his work is used as a foundation and for the further development of different levels and classes of actions. The ALAM framework uses the feedback classification of Rogers (1951) and Shute (2007) to generate formative feedback based on the actions performed by the learner. It also contributes to Experiential Learning Theory developed by Kolb (1984). Over the years, Kolb and Kolb (2012) were more involved in the learning and teaching aspects of the theory and not the assessment of actions; the ALAM framework now addresses the lack of an action-based assessment methodology.

1.5.2. Contribution to Practice

In practice, virtual world developers may use the findings to support the development of action recognition, action-sequence recognition, and feedback-generation technologies. The BEHAVE taxonomy (Fardinpour & Reiners, 2014) contributes to various research areas by providing a standard exhaustive classification of human actions for data tagging, action recognition, and performance recognition. The business sector including mining companies and medical institutes may embed the findings into their current (2D/3D) VTEs to assess their learners (Section 2.4).

1.6. Organisation of the Thesis

The thesis consists of six chapters. Chapter 2 investigates the literature relevant to Action-based Learning, learning assessment, the taxonomy of human actions, virtual environments, and simulation-based training. Action-based Learning and learning assessment are introduced, and various taxonomies and classifications of human actions in various disciplines are examined. Various simulation-based environments are investigated and described, and their differences are pointed out. The research methodology, Design Science Research, research design, and research questions are discussed and illustrated in Chapter 3. Chapter 4 introduces the Action-based Learning Assessment Methodology as well as the conceptual model of an assessment system. In Chapter 5, BEHAVE and its application are explained and illustrated with examples. Each level and class of the taxonomy is defined, and the BEHAVE syntax is applied to different examples. The data and analysis resulting from the evaluation of BEHAVE conducted in Chapter 6, is discussed in Chapter 7, followed by the possibilities for future research, and conclusion in Chapter 8.

Chapter 2: Literature Review

2.1. Introduction

Training and assessment has advanced over time with the ongoing investigation of learning behaviours and the development of new learning theories and technologies. The development of new learning theories and methodologies has changed traditional classroom practices including teaching, assessment, communication, and the learning environment (Figure 2.1). New learning theories have moved towards student-centred approaches and more attention has been given to the application of knowledge rather than the memorisation of information. The lecture-exam communication method has evolved into more interactive methods such as group discussions and online forums. The learning environments have expanded from a classroom-only focus by the use of computers, computer networks, simulation-based environments, and games. New educational technologies have given learners greater opportunities to engage in more interactive learning scenarios. While the teaching aspect of learning methodologies is moving towards learning from experiences, the assessment aspect is also continuing to evolve.

Figure 2.1: Change of learning environment, communication, and teaching (All the images are Public Domain and free to use, share, and modify. https://creativecommons.org/licenses/by/2.0/ and https://creativecommons.org/licenses/by-sa/3.0/)

Students might find a way to overcome the effects of bad teaching, but they certainly cannot escape the results of bad assessment if they want to graduate. Graduation might not be the goal in all teaching scenarios, but in most cases, it is a secondary goal at least. Moreover, without assessment, it would be hard to reflect on the learner’s acquired knowledge. Although assessment was not always popular in educational research and theory (Boud, 1995) with a focus on formative feedback being used for learning, its importance has been re-acknowledged (Knight, 2012; Earl, 2012; Black & Wiliam, 2012; Hargreaves et al., 2014). The term ‘formative assessment’ suggests the use of detailed feedback for learning (Section 2.3.1). With the increasing demand for applicability of the acquired knowledge, the lack of an appropriate assessment method involving the learners' performance as a demonstration of knowledge applicability becomes more evident.

Typical classroom situations require learners to listen to the teacher and undergo a general assessment to show their knowledge. The traditional teaching and assessment methods tend to require the learner to memorise the material or, at best, demonstrate some understanding of it. Hence, the learner might not have the opportunity to apply knowledge. However, in the cognitive domain, memorisation is on a low level of the learning pyramid according to Bloom’s taxonomy (Bloom et al., 1956). Based on Bloom’s taxonomy (Bloom et al., 1956), although the memorising and understanding of knowledge is necessary, it is not sufficient for knowledge application to problem-solving in real-life scenarios. In contrast, Action-based Learning assesses the learners’ knowledge at the application level which will demonstrate that learners can use their knowledge in practice. Action-based Learning (Section 2.1) includes, but is not limited to: Problem-based Learning (Barrows & Tamblyn, 1980); Experiential learning (Kolb, 1984; Kolb & Kolb, 2012); Inquiry or Goal-based Learning (Schwab, 1960; Herron, 1971; Edelson et al., 1999; Wilhelm & Wilhelm, 2010); Adventure Learning (Doering, 2006); and Scenario-based Learning (Naidu, 2010). Action-based Learning theories define a framework for how learners can use an environment to practice what they have learned and apply lessons to real-life problems. Although Action-based Learning has advanced over time (with more than 2500 research publications just on experiential learning theory (Kolb and Kolb 2010 a, b)), its assessment has remained limited. Investigation of Action-based Learning assessment literature shows: a) significant difference in the number of publications on teaching and assessment methods respectively, and b) the use of similar assessment methods over time such as general assessment, peer assessment, written essay, and performance review (Segers & Dochy 2001; Sluijsmans et al., 2001; Gijbels et al., 2005; Teaching and Learning Services, 2014). The current state of the art assessment approach requires the learner’s performance to be observed by an expert in real-time or saved on a video recording. Although expert feedback is likely to be more effective than the learner being assessed by exams and quizzes, the limited temporal, geographical, and financial issues associated with assessment by experts restricts scalability and sustainability, and calls for solutions via the automation of assessment. Thus, automation of Action-based Learning scenarios would allow more learners to benefit from the feedback from highly recognised experts in each field without the direct involvement of these experts. The assessment of learners’ performance during Action-based Learning via detailed formative expert feedback is intended to raise the learning to the evaluation level of Bloom’s taxonomy. This elevation in learning level is due to the detailed information provided to the learner, enabling him/her to evaluate the performance and learn from different reference solutions.

There are numerous skills-based activities which are considered by industry training experts to be too hazardous, expensive, or time-consuming to be practised in real-life settings. Hence, simulated environments, where learners can acquire and practise such skills, are replacing the real-life learning environments. As Eschenbrenner et al. (2008) noted, virtual worlds and new virtual reality technologies, especially 3D VTEs, can improve learning and assessment. Nowadays, virtual worlds are widely used in education for learning support (Duncan et al., 2012). The simulation-based environments allow learners to practise and apply what they have learned. However, even though the environment for practice has improved, the examinations used to test learners' knowledge are usually general assessment methods including short answers, multiple choice questions, or selections (Ong, 2007).

This chapter investigates Action-based Learning and its various theories, and learning assessment and automated assessment, in order to demonstrate the need for an Action-based Learning Assessment Methodology. Due to the focus of Action-based Learning assessment on learner's actions, the chapter investigates various taxonomies and classifications of human actions as well as simulated-based learning and assessment environments.

2.2. Action-based Learning

Experience-based learning theories are often classified under the broad terms 'learning-by-doing' or 'Action-based Learning' (Logan & Stuart, 1987; Bruce & Bloch, 2012; Naidu & Bedgood, 2012). Learning-by-Doing or Action-based Learning is a valuable theory for educators and researchers in education and refers to "all learning that is orchestrated by some activity on the part of learners" (Naidu & Bedgood, 2012). Naidu & Bedgood (2012) state that legitimate learning actions may vary from real participation by learners (in building, creating, or drawing something) (e.g. Figure 2.2) to learners watching a video clip that is later examined, reflected on, or plants a seed for a subsequent decision-making process. Considering that experience and action are the core of various Action-based Learning methods, the use of video seems to be at odds with the characteristics of Action-based Learning.

Figure 2.2: A volunteer from C3KC teaches Sarah to solder (by Wesley Fryer on Flickr - Free to share, license: https://creativecommons.org/licenses/by/2.0/)

When we experience an event, we can learn from that event retrospectively, concurrently, and prospectively, i.e.:

- Learning from the event while it is happening,
- Learning from an event in the past, when thinking about it later,
- Learning more about an event in the past, when thinking more about it,
- Reinterpreting an event from past in a different way, in the light of further experience(s),
- Analysing future scenarios (Beard & Wilson, 2013: pp. 44 - 45).

An important point missing from Beard & Wilson (2013)'s five chronologically listed learning situations is the provision of feedback for learning. In a situation where the learner has made mistakes during the performance, s/he is not able to correct those mistakes by reviewing the performance unless the correct solution is provided through feedback.

While each Action-based Learning type has a distinctive focus or perspective, each starts from a defined problem or learning goal which has to be solved or achieved (Naidu, 2010). Action-based Learning characterises a learner-centric model where the learner studies the learning material and then applies the lesson learned. The 'learning-by-doing' approach distinguishes Action-based Learning from simple 'action learning' where learning is achieved "by using personal experience and reflection, group discussion, and analysis, trial-and-error discovery, and learning from one another" (Lasky & Tempone, 2004: p. 87). Action Learning is evident, for example, within a group of employees who are discussing, analysing and solving particular problems. In contrast, Action-based Learning is about performing in the learning environment in order to achieve a learning outcome. The focus on business problems and scenarios differentiates Action Learning from Action-based Learning. In the following subsections, the different identifying characteristics of a variety of action-based theories are described and discussed.

2.2.1. Experiential Learning

Kolb (1984: p. 41) defines learning as "the process whereby knowledge is created through the transformation of experience. Knowledge results from the combination of grasping and transforming experience". Experiential Learning Theory (ELT) was based on the work of researchers such as William James, John Dewey, Kurt Lewin, Jean Piaget, Lev Vygotsky, Carl Jung, Paulo Freire, Carl Rogers, and others (Figure 2.3), to develop a "dynamic, holistic model of the process of learning from experience and a multi-dimensional model of adult development" (Kolb & Kolb, 2012: p. 1215).

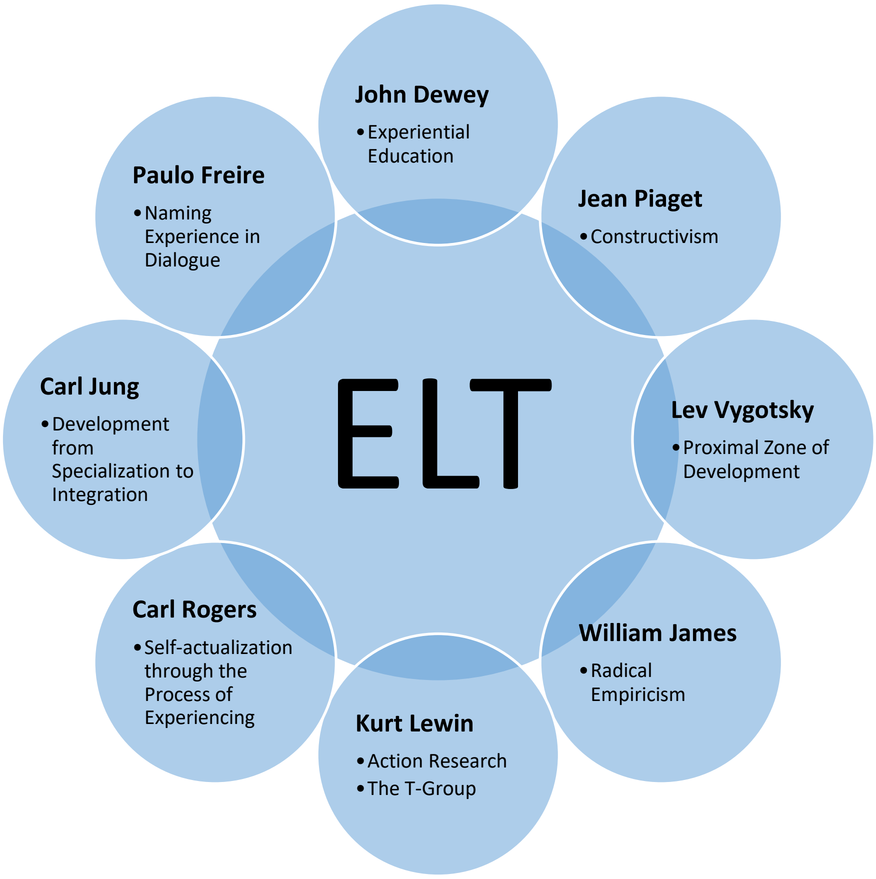


Figure 2.3: Foundational scholars of experiential learning theory (Kolb & Kolb, 2012)

ELT combines the works of the initial experiential learning researchers around six propositions that they all share (Kolb & Kolb, 2012):

1. "Learning is best conceived as a process, not in terms of outcomes" (p. 1216)

Learning occurs in a process and from outcomes, feedback, differences, and new experiences. Knowledge might not show through performance and the learner might need to have several connected experiences.

2. "All learning is re-learning" (p. 1216)

Learning as a process should force the learner to experience and re-think and experience again and learn from the new ideas and thinking.

3. "Learning requires the resolution of conflicts between dialectically opposed modes of adaptation to the world" (p. 1216)

Learning requires conflict and opposition. Learners need to face different opinions and rethink and react, in order to learn from these differences.

4. "Learning is a holistic process of adaptation" (p. 1216)

Learning does not occur only in formal classroom settings. Learners can have different experiences by facing various situations and learning during the process of solving problems or making decisions.

5. "Learning results from synergetic transactions between the person and the environment" (p. 1216)

Learning is affected by both the learner and the learning environment. Current experiences evolve, and the learner acquires new experiences by interacting with the environment.

6. "Learning is the process of creating knowledge" (p. 1216)

Knowledge is created from the learner’s knowledge, and then this process reoccurs, and new knowledge emerges from re-learning from that current knowledge.

A study of these six propositions highlights the importance of: a) the learning environment, b) different opinions, c) reflection on learner’s performance, and d) freedom of interact with the environment. These four essential characteristics are the platform for a new assessment method that provides learners with a detailed formative feedback that enables them to reflect on their performance in relation to multiple expert opinions.

2.2.2. Problem-based Learning

Since the 1950s when Problem-based learning (PBL) was conceived and implemented to improve medical students' unsatisfactory clinical performances, this instructional method has been used to prepare learners for real-world settings. By requiring learners to solve problems, PBL enhances learners' learning outcomes by stimulating their abilities and skills in applying knowledge, solving problems, practising higher order thinking, and self-directing their learning (Jonassen & Hung, 2012). The format and processes of PBL seen today were first developed in the medical school at McMaster University in the 1960s and 1970s (Barrows, 1996).

Hung et al. (2008) characterise PBL as follows:

- being focused on a simulated authentic problem;
- being learner-cantered;
- being self-directed with self-assessment and/or peer assessment;
- being self-reflective;
- tutors acting as facilitators.

According to the above-mentioned characteristics of PBL by Hung et al. (2008), the learner is responsible for organising the learning strategies, and tutors play a guidance role. Moreover, the learner addresses an ill-structured simulated authentic problem in order to learn about the topic. Finally, learning assessment is done by peers or as self-assessment by the learner. The process of PBL proposed by Hung et al. (2008) is shown in Figure 2.4.

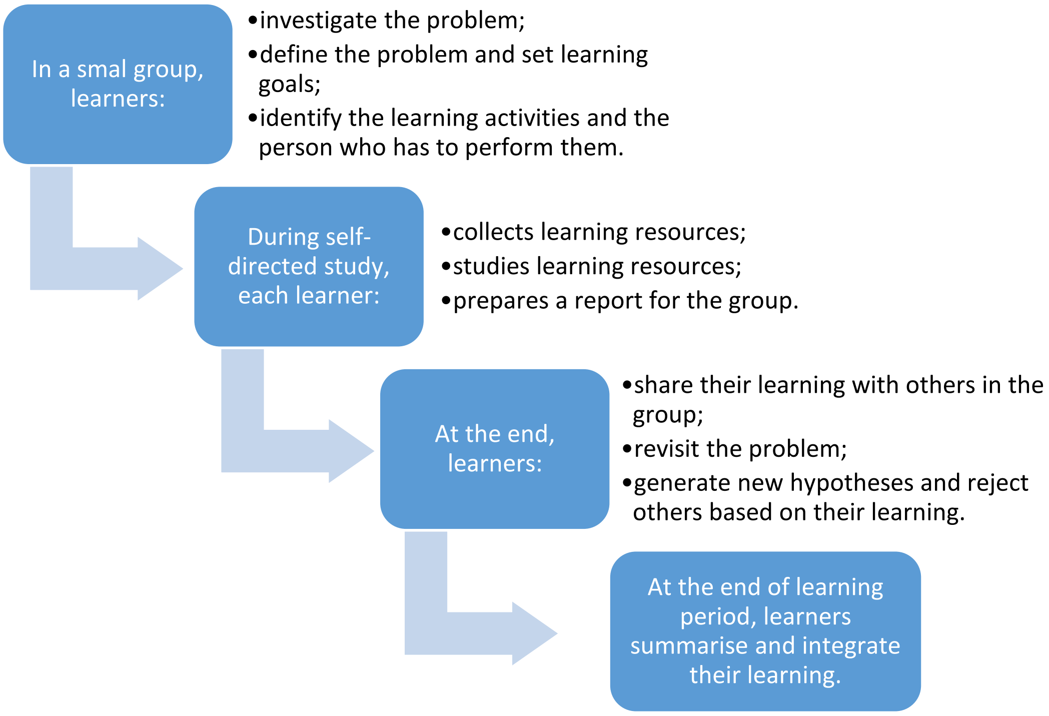

In a smal group, learners:
•investigate the problem;
•define the problem and set learning goals;
•identify the learning activities and the person who has to perform them.
During self-directed study, each learner:
•collects learning resources;
•studies learning resources;
•prepares a report for the group.
At the end, learners:
•share their learning with others in the group;
•revisit the problem;
•generate new hypotheses and reject others based on their learning.
At the end of learning period, learners summarise and integrate their learning.

Figure 2.4: PBL Process by Hung et al. (2008)

2.2.3. Authentic Learning

Stating that the term 'authentic' is applied "loosely and inconsistently to a wide range of theoretical and practical work" (p. 195), Shaffer & Resnick (1999) identify four different types of authentic learning: "(a) learning that is personally meaningful for the learner, (b) learning that relates to the real-world outside of school, (c) learning that provides an opportunity to think in the modes of a particular discipline, and (d) learning where the means of assessment reflect the learning process" (p. 195). Shaffer & Resnick (1999) state that these four types of authentic learning are interdependent; they cannot succeed without each other. Reviewing the current literature of authenticity and authentic learning, Shaffer & Resnick (1999) found four significant meanings for authentic learning with the greatest number of references being in ERIC online library:

a) "materials and activities aligned with the world outside the classroom";
b) "assessment aligned with (what students really should learn from) instruction";
c) "topics of study aligned with what learners want to know";
d) "methods of inquiry aligned with the essential practices of a discipline" (p. 197).

Shaffer & Resnick (1999) considered ‘thick authenticity’ as a vital characteristic of authentic learning environments. Thick authenticity refers to activities “that are personally meaningful, connected to important and interesting aspects of the world beyond the classroom, grounded in a systematic approach to thinking about problems and issues, and which provide for evaluation that is meaningfully related to the topics and methods being studied.” (p. 203). They strongly suggest that computers are the best media for creating a thick, authentic learning environment.

Reeves et al. (2002) identified ten characteristics of authentic activities:

1. “Authentic activities have real-world relevance”;
2. “Authentic activities are ill-defined, requiring students to define the tasks and sub-tasks needed to complete the activity”;
3. “Authentic activities comprise complex tasks to be investigated by students over a sustained period of time”;
4. “Authentic activities provide the opportunity for students to examine the task from different perspectives, using a variety of resources”;
5. “Authentic activities provide the opportunity to collaborate”;
6. “Authentic activities provide the opportunity to reflect”;
7. “Authentic activities can be integrated and applied across different subject areas and lead beyond domain-specific outcomes”;
8. “Authentic activities are seamlessly integrated with assessment”;
9. “Authentic activities create polished products valuable in their own right rather than as preparation for something else”;
10. “Authentic activities allow competing solutions and diversity of outcome” (p. 564).

Gulikers et al. (2005) differentiated an authentic learning environment from an authentic task. Authentic environments provide “a realistic context to an authentic task”, that is, a learning task resembling a “task performed in a non-educational setting and that requires students to apply a broad range of knowledge and skills” (p. 510).

Authentic learning occurs when learners learn to solve real-world problems by performing authentic activities such as role-playing exercises, problem-based activities, and case studies, in an authentic learning environment and benefit from feedback by learning from their performance.

2.2.4. Inquiry-based Learning

Enquiry/Inquiry: “The action of seeking… truth, knowledge, or information concerning something; search, research, investigation, examination” (OED inquiry/enquiry 1.a)

Inquiry-based learning refers to a learning process that requires learners to search for knowledge by questioning and exploring the problems. A learning process usually starts with the formulation of a question or a problem. Then learners try to create plans to examine, cooperate, justify, and reflect on the solution for the problem or answer to the question. They later communicate and share the conclusion (Caliskan, 2012). Inquiry takes various forms, including: an inquiry and design framework; understanding by design; project-based learning; problem-based learning; and case-based learning (Wilhelm & Wilhelm, 2010). Due to the nature of this learning method, learners do not passively receive the knowledge; instead, they actively try to understand and use and share it in real-life problem solving. Inquiry-based learning is very similar to PBL, the only difference between the two being the role of the tutor. While in Inquiry-based learning the tutor provides the information and facilitates the learning (encouraging/expecting higher-order thinking), in PBL the tutor supports the learning process, but gathering of information and logical thinking is the learner’s responsibility (Savery, 2006).

2.2.5. Scenario-based Learning

The scenario-based learning premises that in a well-designed learning scenario, in the form of a story, learners play a crucial role of the kind that they might have in real life in the future. Thus, learners are engaged in an authentic learning scenario, where they have to solve a given problem (Iverson & Colky, 2004).

Scenario-based learning (SBL) uses collaborative scenarios to support active learning strategies to resolve problem-based tasks. It involves learners working their way through a scenario, based around a life-like problem, which they are required to solve. In the process, learners must apply their subject knowledge, critical thinking, and problem-solving skills in a safe, authentic context. SBL can provide several feedback opportunities to learners, based on the decisions they make at each stage of the process. SBL might be considered as an authentic PBL, as learning scenarios are presented in an authentic simulated environment (Carroll, 2000; Errington, 2003; Kneebone et al., 2005; Breakey et al., 2008; Rosson & Carroll, 2009).

2.2.6. Adventure Learning

Veletsianos & Kleanthous (2009) introduced various online learning programs that are based on adventure and expedition, such as:

- GoNorth! (http://www.polarhusky.com/);
- The Jason Project (http://www.jason.org);
- The World of Wonders (http://www.questconnect.org/world_of_wonders.htm);
- Blue Zones (http://www.bluezones.com/education);
- Expedschools (http://www.expedschools.org/);
- eField Trips (http://www.efieldtrips.org/).

Doering (2006: p. 200) defined Adventure Learning as "a hybrid[1] online educational environment that provides learners with opportunities to explore real-world issues through authentic learning experiences within collaborative online learning environments". He proposed a framework that consists of seven principles, which are interdependent: 1. a curriculum developed based on problem-solving; 2. learners have the opportunity to collaborate with experts, peers, and interact with the content; 3. the curriculum and learning environment are utilised by the Internet; 4. the curriculum is enriched by the use of various media from the field; 5. active learning curriculum and learning opportunities are concordant, 6. the curriculum and online learning environment using a pedagogical guideline; and 7. the education is based on adventure (Doering, 2006). As Veletsianos & Kleanthous (2009) showed in their review of the various learning approaches, different alternative terms can be used, including adventure learning; virtual/electronic field trips; adventure-education; outdoor education; and online expeditions. Although all these approaches are adventure-based, they vary from educational design models to projects, and from

[1] "AL utilizes both F2F and online learning environments but is subtly different from traditional hybrid environments. For example, in AL environments, classroom teachers are not positioned in the role of teacher/facilitator/designer in the online learning spaces. AL online spaces are collaborative spaces where students, teachers, experts, and AL team members interact with each other; these are community spaces where traditional hierarchical classroom roles are blurred and learning is transformed" (Doering, 2006: p. 198).

virtual environments to outdoors or a mix of the two (Veletsianos & Kleanthous, 2009). Adventure-based learning is based on the theoretical foundations of experiential and inquiry-based learning in an authentic environment (Doering, 2006; Veletsianos & Kleanthous, 2009; Moos & Honkomp, 2011). "More specifically, the approach assumes that students learn by immersing themselves in participatory experiences grounded in inquiry" (Veletsianos & Kleanthous, 2009: p. 86).

2.3. Learning Assessment

The terms ‘assessment’ and ‘evaluation’ have been used inversely in different educational systems. For example, in the UK, assessment refers to the judgement made about the learner’s work, while evaluation refers to the judgement of courses, course delivery, and/or the process of making such judgements; however, this is the opposite in the USA (Taras, 2005). In this research evaluation is “a process of measuring the quality of a performance, work product or use of a process against a set of standards to make a judgment or determination if, or to what level, the standards have been met” (Baehr, 2005: p. 1), while assessment is “a process of measuring and analysing a performance or product to provide feedback to improve future performance or products” (Baehr, 2005: p. 1). Assessment has four components: setting up, designing, performing, and reporting the assessment (Baehr, 2005). The focus of this research is on learning assessment and not the evaluation.

2.3.1. Summative and Formative Assessment

Learning assessment is categorised into the Summative, Diagnostic, and Formative Assessment (McTighe & O’Connor, 2005). Diagnostic assessment, or pre-assessment, typically occurs before the commencement of instruction. It is used to: check the learner's prior knowledge and skills, identify the learner’s misunderstandings, outline the learner's interests, and discover the learner’s preferred learning-style. Diagnostic assessments provide information to assist the educator to plan and guide learners individually. Different types of diagnostic assessments include checking the prior knowledge and skills, and interest or learning preference surveys and examinations. Because pre-assessments are used for analytical purposes, educators normally do not grade the results (McTighe & O’Connor, 2005).

Michael Scriven coined the terms ‘formative’ and ‘summative’ in 1967, developing two of the earliest definitions which are still in common use today. That is, formative assessment is used when the purpose is learner improvement and summative assessment is the judgement that will result in pass/fail decisions (Scriven, 1967). Bloom also uses the same terminologies and definitions. According to Bloom, formative assessment is used "to provide feedback and correctives at each stage in the teaching-learning process" (Bloom, 1969: p. 48), while summative assessment is used to judge what the learner has achieved at the end of a course or curriculum (Bloom, 1969). Black & Wiliam (2009) defined formative assessment as “practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited” (p. 9). Taras (2005) defines summative assessment as “the process of assessment leads to summative assessment, that is, a judgement which encapsulates all the evidence up to a given point. This point is seen as a finality at the point of the judgement. A summative assessment can have various functions which do not impinge on the process” (p. 468). Although these definitions were presented more than three decades after Scriven coined the terms ‘summative’ and ‘formative’, the core of the definitions remains the same. The following paragraphs shed more light on summative and formative assessment.

Summative assessment differs from formative assessment in that it is more focused on tallying or summarising the accomplishments of learners, and is focused on reporting at the end of a course of study, especially for the purposes of certification (Sadler, 1989). Although Sadler and Scriven’s definitions are still relevant and highly cited today, various formats and methods of summative assessment have been developed in educational institutes including oral presentation (Turner et al., 2013), short-answer quiz, multiple-choice quiz, reading exams (McDaniel et al., 2012), and written tests such as reports and essays. Investigating the importance of multiple-choice tests in education in recent years, Roberts (2006) repeatedly cited numerous studies showing the extensive use of this assessment method as an exclusive summative assessment. He listed various advantages of this type of assessment such as quick testing of knowledge in significant groups, being automatically marked, and re-usability. However, a number of disadvantages are also mentioned. For example, it tests recall only, takes a long time to construct, and is not well suited to test creativity or unique thinking (Roberts, 2006). Both Cilliers et al. (2010) and Trotter (2006) investigated the effect of summative assessment that is too frequent, on students' learning behaviour and their results strongly suggest that a constant repeated summative assessment of students changes the students' learning patterns.

Formative assessment or 'assessment for learning' is considered by some researchers and educators as "… not a test but a process" (Popham, 2008: p. 6). Black & Wiliam (2009: p. 8) named five strategies comprising the concept of formative assessment:

1. "Clarifying and sharing learning intentions and criteria for success";
2. "Engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding";
3. "Providing feedback that moves learners forward";
4. "Activating students as instructional resources for one another; and";
5. "Activating students as the owners of their own learning".

In a study of the literature pertaining to online formative feedback in higher education, Gikandi et al. (2011) reported that Scenario-based learning, Collaborative learning, Authentic learning, Communities of Practice (COP), Problem-based learning, and Active learning are the most cited learning theories in using formative assessment. The majority of these theories are considered as a sort of Action-based Learning. Gikandi et al. (2011: p. 2341) pointed out that "online formative assessment can function as an innovative pedagogical strategy through facilitating the following opportunities: (1) formative and immediate feedback, (2) engagement with critical learning processes, and (3) promoting equitable education". These opportunities can be actualised using different technologically-facilitated learning methods.

Clark (2012) presented the results of his investigation of 199 research publications on formative assessment and its role in Self-Regulated Learning (SRL), concluding that formative assessment actualises and reinforces SRL strategies among students with the help of formative feedback, synchronous feedback, and external/internal feedback. The formative feedback improves the involvement of pupils in "meta-cognitive strategies such as personal goal-planning, monitoring, and reflection" (p. 210). Activities which result in an immediate feedback are very engaging, and that is why "Synchronous feedback has been found to enhance learning and be more effective at supporting higher psychological functioning, such as synthesis" (Clark, 2012: p.210). Self-regulated learners generate more internal feedback, respond positively to external feedback, and put more effort into achieving their learning goals compared with non-self-regulated learners (Clark, 2012). According to Clark, theoreticians agree that, because learners gain from the adaptive and self-directed learning, SRL seems to improve academic outcomes and motivation in learners. Clark (2012) found the theory of formative assessment to be a "unifying theory of instruction, which guides practice and improves the learning process by developing SRL strategies among learners" (p. 205).

Formative assessment also contributes to the improvement of learning attitudes in mobile learning. Hwang and Chang (2011) introduced a mobile learning system using Formative Assessment-based Mobile Learning (FAML) mechanism, which guides learners to observe real-world learning objects and interact with them. After the answer has been submitted, instead of the correct answer, FAML guides learners to find the correct answers on their own by giving them hints and supplementary materials based on their answers.

Formative assessment is also used in serious games for learning or/and game-based learning. It can be both executed by the teacher or be embedded in the game. Belland (2012) named the different ways that formative feedback can be provided to learners using educational games. They vary from rubrics used by the teacher and rubrics sent out to students for peer assessment, to written feedback and short debriefing sessions. However, when it comes to embedding the formative assessment into the games, the main concern seems to be detracting from the learner's interest (Walker & Shelton, 2008). To overcome this dilemma, Shute (2013) suggested the use of stealth assessment in educational games, which are "… assessments that are woven directly and invisibly into the fabric of the gaming environment" (p. 29), so learners would be assessed based on their interactions with the game without being interrupted. Stealth Assessment is based on the choice of actions and action sequences in games for learning. Al-Samadi et al. (2010) proposed a framework for the design of assessment and feedback for serious games using Stealth Assessment. Their approach is to assess action choices in games, and create formative feedback at an 'interpretive' level of Rogers' feedback classification (Rogers, 1951), whereby players receive a score and have their wrong action highlighted so they can adjust.

In all the different applications of formative assessment discussed above, formative feedback is the core activity regardless of the technologies being used, or the learning theories that are investigated in the next section.

2.3.2. Feedback

Feedback comprises the interpretation of the learners' assessment based on the provided learning material and activities to invoke a behavioural change in future learning activities (Black & Wiliam, 1998). Feedback is a crucial part of formative assessment and is usually defined "in terms of information about how successfully something has been or is being done" (Sadler, 1989: p. 120). The research presented in this thesis adopts the feedback definition of Hattie and Timperley (2007: p. 81): "feedback is conceptualized as information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding". Hattie and Timperley (2007: p. 81) also mentioned that different so-called agents provide different information as feedback in the learning process. For example, "corrective information" by teachers or parents, "alternative strategies" by peers, and books "can provide information to clarify ideas".

Shute (2007: p. 2) noted the various characteristics of formative feedback including: "multidimensional; non-evaluative; supportive; timely; specific; credible; infrequent; and genuine". Black & Wiliam (1998) named two key features of feedback: directive and facilitative. Directive feedback provides learners with what needs to be fixed or revised, while facilitative feedback guides learners in their revision and conceptualization. Formative feedback may be used by learners via three mechanisms (Black & Wiliam, 1998: p.157):

1. "it can signal a gap between a current level of performance and some desired level of performance or goal";
2. "it can effectively reduce the cognitive load of a learner, especially novice or struggling students";
3. "it can provide information that may be useful for correcting inappropriate task strategies, procedural errors, or misconceptions".

Formative feedback information can be provided by different means: "verification of response accuracy, explanation of the correct answer, hints, worked examples" (Shute, 2007: p. i). Shute (2007: p. 10) described 12 different types of feedback based on their complexity (See Table 2.1), starting from the lowest level of

‘No feedback’ to the highest of ‘Informative tutoring’. This 12-class feedback classification relates to Roger’s (1951) very first classification for student-centred learning which was:

1. Evaluative (learners receive a score)
2. Interpretive (learners receive a score and the wrong action)
3. Supportive (students receive a score and guidance information)
4. Probing (students receive a score and analysis of why the student did the wrong action)
5. Understanding (students receive a score and analysis of why the student performed the wrong action as well as guidance for supportive steps or learning material)

Both of the Shute and Rogers classifications include scores, which can become a point of contention as most experts believe that formative assessment should not be graded. As McTighe and O'Connor (2005: p. 10) stated, "although teachers may record the results of formative assessments, we should not factor these results into summative evaluation and grading".

In this research, the Rogers (1951) and Shute (2007) classifications of different types of feedback are adapted for use as part of the ALAM framework (Section 4.3.3).

Table 2.1: Feedback Types Arrayed Loosely by Complexity (Shute, 2007)

Feedback type	Description
No feedback	Learners do not receive any results.
Verification	Learners receive the overall mark or at most the right/wrong result of the answers.
Correct response	Learners receive only the correct answers.
Try-again	Learners receive the incorrect response and are allowed to answer the question again until they get the correct answer.
Error-flagging	Learners receive the highlighted errors, without the correct answer.
Elaborated	Learners receive an explanation of why an answer is right or wrong, and they might be able to review the instruction or the correct answer (see below for six types of elaborated feedback).
Attribute isolation	An elaborated feedback that provides information on crucial features of the studied concept or skill.
Topic-contingent	An elaborated feedback that provides information related to the studied topic such as re-teaching content.
Response-contingent	An elaborated feedback that provides information on a specific answer explaining why the answer is wrong or right.
Hints/cues/prompts	An elaborated feedback that provides guidance on the right direction (e.g., hints, examples, or demonstrations); but without the correct answer.
Bugs/misconceptions	An elaborated feedback that provides information about explicit errors or misunderstandings.
Informative tutoring	An elaborated feedback that provides verification feedback, error-flagging, and deliberate hints on how to continue; but usually without the correct answer.

2.3.3. Authentic Assessment

In contemporary literature, the term ‘authentic’ is often aligned with having something to do with the ‘real world’ (Barab et al., 2000; Herrington & Herrington, 2008; Frey et al., 2012). Although there is no consensus among researchers regarding the definition of authentic assessment, one unique and widely-accepted characteristic of authenticity is its realistic nature (Frey et al., 2012). Gulikers et al. (2004) defined authentic assessment as: “an assessment requiring students to use the same competencies, or combinations of knowledge, skills, and attitudes that they need to apply in the criterion situation in professional life” (p. 69). Swaffield (2011) used the same criteria, real-world-like tasks and meaningful situations, to define authentic assessment as one that is “conducted through ‘real world’ tasks requiring students to demonstrate their knowledge and skills in meaningful contexts” (p. 434).

In common with other theories, authentic assessment has been subjected to the interchangeable use of concepts, and the term has sometimes been used interchangeably with performance assessment and formative assessment (Hart, 1994; Torrance, 1995). Researchers such as Herrington & Herrington (1998), Gulikers et al.

(2004) and Frey & Schmitt (2007) have attempted to provide more objective criteria to differentiate between these terms. Authentic assessment can be compared with the traditional assessment methods to enable a better understanding of ‘authenticity’ (Herrington & Herrington, 1998; Wiggins, 1990). Table 2.2 compares authentic assessment with the traditional assessment.

Table 2.2: A comparison of authentic and traditional assessment (Herrington & Herrington, 1998; Wiggins, 1990)

Authentic assessment	Traditional assessment
The performance is directly assessed based on educated rational tasks	It is based on tools reflecting the memorisation of the knowledge
Encourages effective performance by the use of learned knowledge	Shows the recognition and recall of the learned knowledge
Full range of tasks	The usual pencil-and-paper and one-answer questions
Refined, in-depth and reasonable answers, performances or products are expected	Only the correct responses are expected, sometimes regardless of explanations
Scoring attains validity and reliability by highlighting and standardising the suitable criteria	Objective ‘items’ and a ‘right’ answer for each is the standard practice
It is valid if it simulates real- world assessment of skills	It is valid if it matches with the curriculum content or/and there is a correlation with other test results
It is based on an ill-structured problem	It assesses random disconnected or simplistic elements of activities

Gulikers et al. (2004) developed a five-dimensional framework for authentic assessment and stated that if these five dimensions are designed to be authentic, this will ensure the authenticity of the assessment. These five dimensions are: "(a) the assessment task, (b) the physical context, (c) the social context, (d) the assessment result or form, and (e) the assessment criteria" (p. 70).

Frey et al. (2012) identified various characteristics and components of authentic assessment. These include realistic context or activity, performance-based activity, cognitive complexity, the answer needs a defence, the assessment is formative, the activity is collaborative, the scoring criteria are known or developed by students, scoring is based on multiple indicators, and students are expected to perform with mastery.

Ashford-Rowe et al. (2014) discussed ‘eight critical elements of authentic assessment’ from the literature.

1. “An authentic assessment should be challenging.”

Learners should be challenged just as they would be in real-life scenarios during the assessment. They need to analyse the task and, based on their learned skills, choose the appropriate response.

2. “The outcome of an authentic assessment should be in the form of a performance or product (outcome)”

During the assessment, the learner should demonstrate skill by applying the acquired knowledge. The outcome would be in the form of a product or performance that is judged by the designer as being, or not being, satisfactorily completed.

3. “Authentic assessment design should ensure transfer of knowledge.”

Authentic assessment should ensure that the learner can apply the learned knowledge in other domains as well.

4. “Metacognition as a component of authentic assessment.”

Authentic assessment should enable learners to learn from their assessment. Constructive feedback can help learners to reflect on their knowledge and skills.

5. “The importance of a requirement to ensure accuracy in assessment performance.”

The purpose of authentic assessment is to show both the achievement of a goal and the process involved in that achievement. The learner has to demonstrate the application of the knowledge, not only the understanding of it. Furthermore, authentic assessment has to assess how closely the skills approximate those required for real-life applications.

6. “The role of the assessment environment and the tools used to deliver the assessment task.”

Authentic assessment should provide a closely similar assessment environment for the learner, although sometimes the re-creation of the real world is difficult in a training environment. Culturally familiar settings, such as language and images, also need to be in place.

7. “The importance of formally designing in an opportunity to discuss and provide feedback.”

As discussed above, the learner needs to reflect on his/her performance. Authentic assessment has to provide feedback on the learner's performance so that s/he can learn from mistakes and improve the required skills.

8. "The value of collaboration."

Authentic assessment provides the learner with the opportunity to communicate with others during the assessment process. Collaboration with team members can be used for problem-solving, peer assessment, and information-sharing similar to real-life situations.

Abrams and Grebers (2013) used the feedback loop and formative assessment as an important component of authentic assessment in the use of video games in school assessments. They consider video games as an authentic assessment not because "success hinges on students' application of learned information in a new yet relevant context, but also because students are asked to be creators and/or problem solvers" (p. 96).

Reiners et al. (2014) used authentic assessment in their 'fear of dying' experiment in a virtual environment for training. Their experiment included several components of authentic assessment in that it was performance-based, immersive and realistic, with a complex scenario to be assessed by multiple indicators.

2.3.4. Automated Assessment

Due to the considerable increase in the number of students and consequently the marking loads, educators might be forced to acquiesce to "working more hours (working harder), working more efficiently (working smarter), or they can lower the quality of work (working quickly, marking less accurately and/or with less formative feedback)" (Dreher et al., 2011). With the explosive growth in Internet usage for various purposes (e.g. academic, commercial, personal and educational), education providers are closer to achieving "the concept of offering an education to anyone anyplace at any time" (Hu & Xia, 2010: p. 250). Assessment plays an important role in distance and online education (Hu & Xia, 2010). Automated assessment can empower educators to perform better and smarter when managing resources; moreover, this is not exclusive to online education (Dreher et al., 2011). In the following, various automated assessment systems exemplify the advancement of automated assessment over time.

- The BOSS Online Submission System: is a course management tool, developed by the department of computer science at the University of Warwick. Joy and Luck (1998) started their research on automated assessment in 1993 and continued developing it to the present. BOSS was written in response to increasing class sizes and the need to mark and manage formative assessment for a programming course, to enable the learners to receive feedback on their assignments quickly, efficiently and accurately. The software now includes a completely redesigned automatic testing system (with JUnit[1] testing included), code metrics capabilities and an upgraded, fully integrated Sherlock plagiarism detection system. Version 3.0 was released in September 2003 (Joy et al., 2005) and has been used ever since.

- Browne (2002) introduced an automated assignment assessment system, which aimed to enable engineering students to create computerised mathematical models. It used learner identification numbers as input parameters for the simulations, and the assessment was automated using automated "reverse-engineering" of incorrect answers to find out the source of the error.

- Williams and Dreher (2004) reported on their automated essay grading system, MarkIT, which enables the marking of a vast number of essays using model answers. A combination of natural language processing and pattern matching techniques is used to build a consistent proprietary knowledge representation, and determine the number of the model answer(s) knowledge that is contained in the learner's answer, and a grade is assigned accordingly. To extract lexical information for the building of the document knowledge representation, an electronic version of Roget's Thesaurus was used.

- PASS is another assignment assessment system, which was developed for computer programming courses at the University of Hong Kong (Yu, Poon, & Choy, 2006).

[1] JUnit is a simple framework to write repeatable tests. It is an instance of the xUnit architecture for unit testing frameworks. http://www.junit.org

- MathPASS (Su et al., 2010) and MASS (Modular Assessment System for Modern Learning Settings) (AL-Smadi et al., 2010) were two other assignment assessment systems developed in 2010.

To date, essay grading, automated grading of writing, and assignment grading have continued to be the focus of automated assessment in a number of research projects including: Shen et al. (2001); Palmer et al. (2002); Valenti et al. (2003); Shermis et al. (2010); Salmela & Tarhio (2004); Kopainsky et al. (2012); Toranj & Ansari (2012); Vujošević-Janičić et al. (2013); Foltz et al. (2014a); Foltz et al. (2014b); and Schramma & Srinivasan (2015).

All the automated assessment systems introduced above have been developed for the assessment of essays and short answers; however, none of them assesses human performances. The lack of an automated performance assessment system is an important motive for the development of the Action-based Learning Assessment Method. The conceptual model of this system and a framework for this assessment method are developed and introduced in Chapter 4.

2.4. Virtual Environments for Learning and Assessment

With the development of collaborative learning theory and other similar theories in education, the use of virtual reality and virtual worlds increased rapidly in education. Wilson (1996: p. 8) defined Virtual Learning Environments (VLE) as "computer-based environments that are relatively open systems, allowing interactions and encounters with other participants" and providing access to a wide range of learning resources. Nonetheless, VLE cannot be referred to every educational website, or (3D) VR or a virtual (online) campus (Dillenbourg et al., 2002). A VLE has seven identifying features (Dillenbourg et al., 2002):

1. VLE is a “designed information space”;
2. VLE is a “social space” which the educational interactions turn the “space” into “place”;
3. VLE is “explicitly represented”, varying from “text to 3D immersive worlds”;
4. The learners are co-constructors of the VLE;
5. VLE is not limited to distance education. It can be used in classrooms as well;

6. VLE uses diverse technologies and different educational approaches;
7. VLE converges with physical environments.

Although the use of non-simulation-based VLEs, like content management systems and intelligent tutoring systems, is widespread these days, due to the nature of this research, simulation-based VLE is the centre of interest in this thesis. Simulation-based VLEs use different technologies and environments such as Virtual Reality (VR), Virtual Worlds (VW), and 3D Virtual Training Environments (3D VTE).

In this research, VR is: a set of tools, technologies, and frameworks (Steuer, 1992) which enables a person to create a simulated reality to explore and interact with; such reality in which the person can choose and create the desired time, place and circumstance which is not achievable under real-world conditions (Fuchs & Guitton, 2011).

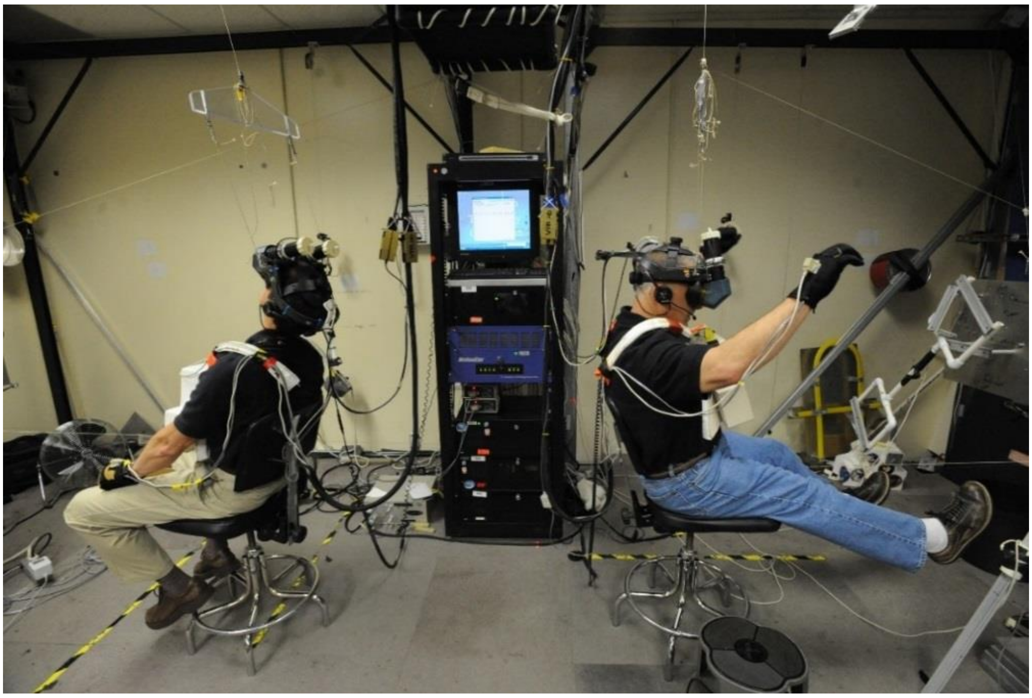


Figure 2.5: Spacewalk in the Integrated EVA/RMS Virtual Reality Simulator Facility at Johnson Space Centre.[1]

VW is an environment that uses VR to connect users to an uninterrupted simulated environment (Figure 2.5) in which users "create a virtual identity which persists beyond the initial session" (Farley, 2011: p. 194). This research used VWs "as

[1] 090324-N-2959L-143 HOUSTON (March 24, 2009) U.S. Navy photo by Mass Communication Specialist 2nd Class Dominique M. Lasco/Released

a computer-based, immersive, 3D multi-user environment that simulates real (or imaginary) life, experienced through a graphical representation of the user" (Gregory et al., 2013: p. 314).

VWs such as Second Life have contributed to different fields including medicine, law, commerce, social life, entertainment, education, and gaming. Learning in virtual worlds transfers to the real world and learners can improve their abilities by learning in virtual worlds (Jarmon et al., 2009). Teaching and training are different approaches to knowledge delivery in any educational setting. Teaching focuses on delivering information and lessons to learners, enabling them to do something themselves (Teach, 2016)[1]. This type of training through practice teaches learners the skills needed for a certain job or activity (Train, 2016)[2]. 3D VLEs used specifically for training are known as Virtual Training Environments (VTEs). It is common training practice for various industries nowadays to use 3D simulation-based training systems.

Virtual Environments for Training (VET) or Virtual Training Environments (VTE) create a virtual representation of real-world training scenarios. The main goals of using VTE for education are to train learners to operate complex machinery, to respond properly to quickly unfolding events (such as combat decisions), or to be trained in environments that in real-world situations are too expensive or hostile (Moskaliuk et al., 2013).

Usually, in real-world training situations, learners do not have the opportunity to repeat parts of the training until they master particular concepts, principles, or skills; the use of VTE solves this problem (Tichon, 2007) by enabling learners to repeat the activities for as long as it takes to master the concepts, principles and skills - one of the important benefits of VTEs.

Learners can review their performance at every stage in order to reflect on their performance. In addition, most VTEs offer the opportunity for some debriefing and

[1] Teach (Teach, 2016):
[Def. 1] to give lessons to students in a school, college, university, etc.; to help somebody learn something by giving information about it
[Def. 2] to show somebody how to do something so that they will be able to do it themselves

[2] Train (Train, 2016)
[Def. 1] to teach a person or an animal the skills for a particular job or activity
[Def. 2] to prepare yourself/somebody for a particular activity, especially a sport, by doing a lot of exercise

feedback (Moskaliuk et al., 2013; Tichon, 2007), although the provided debriefing is in most cases limited to the “replay of training scenes or a change of perspective” (Moskaliuk et al., 2013: p. 195). Some of the benefits of using VTE in education include personalization, active learning, experiential learning, learner-cantered learning and immediate feedback (Bogdanovych et al., 2009).

This research defines VTE as a training-specific VLE that creates a safe, immersive, authentic, and accessible training environment in which to practise, re-assess, and master skills.

Various VTEs have been developed during the last two decades for the training of pilots, fire fighters, drivers, divers and console operators (for an overview, see Rose et al. 2000). The need for VTEs is well-established due to “lower costs and risks, no need for available equipment to train on and control of pedagogical situations”, which are just a few examples of the numerous benefits of VTEs (Gerbaud, Mollet, & Arnaldi, 2007). Different industries are using virtual training systems to train their employees. Surgical institutes, mining companies, nuclear power plants, and manufacturers are reducing costs and hazards by using VR, VW, and VTEs.

VR, VW, and VTEs have been used to train and assess learners in different fields as an interactive life-like learning environment from the very first training VR systems. Table 2.3 summarises some VTEs.

Table 2.3: Early developed VTEs

VTE	Application	Literature
DIVE	A multi-user VR platform for training	Carlsson & Hagsand, 1993
STEVE	An autonomous, animated tutor that cohabits the virtual world with students. He instructs and helps students learn to perform physical, procedural tasks, such as operating and repairing equipment.	Rickel and Johnson, 1998, 1999
No Name	A simulation-based, problem-solving assessment system for dental hygienists	Mislevy et al., 1999
PRVIR	An integrated VR-based intelligent tutoring system (ITS) which was used to train staff of Nuclear Power Plants for radiological protection	Méndez et al., 2001
JACK	An integration of STEVE with a new VW named HeSPI. HeSPI was developed as a tool for equipment operation and maintenance training for Nuclear Power Plants.	Méndez et al., 2003, 2004
MASCARET	Using "multi-agents systems to simulate realistic, collaborative and adaptive environments for training. This model aims at organizing the interactions between agents and provides them abilities to evolve in this context. In addition, it allows the establishment of models necessary to the creation of Intelligent Tutoring system. MASCARET also permits to define pedagogical activities" (p. 423).	Buche et al., 2003
SECUREVI	To train fire fighter officers in operational management and command	Buche et al., 2003
No Name	Scenario-sharing in a collaborative VTE, applied to GVT (Generic Virtual Training) project.	Gerbaud & Arnaldi, 2008
No Name	Haptics-based virtual environment system for assembly training of complex products with physics-based modelling and haptics feedback.	Xia et al., 2012
No Name	A virtual reality interactive training environment prototype for construction industry offsite production	Goulding et al., 2012
VR-based Fire Training Simulator	A visualization technique based on volume rendering and fire dynamics data to create a realistic and accurate smoke environment. An integrated assessment model of smoke hazards to assess the safety of different paths for evacuation or rescue in virtual training.	Xu et al., 2014

Medical science benefits from 3D simulation training such as haptic devices, VTEs for dentistry and surgery training, primarily to assess medical students' performances. Heinrichs et al. (2008) used virtual worlds to train hospital staff teams in acute-care medicine. They used virtual worlds for: training emergency department (ED) teams to manage individual trauma cases; pre-hospital and in-hospital disaster preparedness training; and training ED and hospital staff to manage mass casualties after chemical, biological, radiological, nuclear, or explosive incidents. They created realistic virtual victims of trauma (6 cases), nerve toxin exposure (10 cases), and blast trauma (10 cases). The latter two groups were supported by rules-based, pathophysiologic models of asphyxia and hypovolemia. Results obtained by Panait et al. (2011) indicated an improvement in the laparoscopic skills of their surgical residents, using the LapSim curriculum. Van Sickle et al. (2011) reported significant improvements in the Gastrointestinal Endoscopic Skills of their residents at the Texas Association of Surgical Skills Laboratories (TASSL) where a flexible endoscopy curriculum was offered. The Computer-Assisted Rehabilitation Program – Virtual Reality (CARP-VR) was developed and used by Dores et al. (2012) in the area of rehabilitation of EF, and also in other cognitive functions such as visuospatial functions, attention and memory, in patients with ABI. Investigations into the use of VR, VTE, and haptics in the medical field have produced a long list of research that is beyond the scope of this research.

There are different assessment systems in VTEs mainly in, but not limited to, the medical field, using Bayesian Networks and Markov Model, such as that of Moraes et al. (2012). Although statistical methods seem to be used more frequently in assessment in VTEs, other methods such as haptics feedback and mathematical algorithms are also used. For example, Perrenot et al. (2012) studied the validity of DV-Trainer (MIMIC Technologies), an assessment tool for robotic surgical skills. DV-Trainer is a robotic surgery simulator that uses a scoring utility with seven criteria: time, the economy of motion, drops, instrument collisions, excessive instrument force, instruments out of view, and master workspace range. A total percentage score representing a combination of these criteria is automatically generated by a computerised algorithm created by the manufacturer.

Although the development of different assessment systems in 3D virtual environments for different fields of knowledge still continues, different researchers believe that it is equally important to improve upon the current assessment methods. Shute (2007) investigated the effect of interruptive feedback appearances during serious games for learning. She proposed a Stealth Assessment method in which the assessment and feedback do not interrupt the performance. Shute et al. (2009, p. 299) defined Stealth Assessment as: “When embedded assessments are so seamlessly woven into the fabric of the learning environment that they are virtually invisible, we call this stealth assessment. Such stealth assessment can be accomplished via automated scoring and machine-based reasoning techniques to infer things that would be too hard for humans (e.g. estimating the value of evidence-based competencies across a network of skills)”. Stealth Assessment is based on the choice of actions and action-sequences in serious games for learning. Al-Samadi et al. (2010) proposed a framework using stealth assessment to assess action choices and sequences in serious games for learning, creating formative feedback on the interpretive level of Rogers' feedback classification (Rogers, 1951). Nelson et al. (2014) investigated the "impact of visual signalling techniques used in a virtual world-based assessment of science inquiry and content on (1) student cognitive load and (2) assessment efficiency" (p. 32). The signalling principle in designing assessment in immersive virtual environments states that learning improves through the use of visual or auditory cues that call attention to the organisation of important material to be learned. Their study showed a significant increase in the efficiency of assessment.

2.5. Taxonomy of Human Actions

Human actions have been the focus of studies by researchers from a variety of fields including psychology, human behaviours, industrial engineering, and computer science; however, the focus and terminology of the research were not always consistent. Research on human actions, activities, performance, behaviour, and tasks have been conducted since the early 1960s. The earliest theories of human actions appeared in the literature from the 1960s to the late 1980s (e.g. Willis, 1961; Berliner et al., 1964; Reed, 1967; Oller, 1968; Goldman, 1970; Fleishman et al., 1984). Thereafter, numerous researchers used these theories in studies of human performance and behaviour, including but not limited to Robertson, 1997, 2000; Cappella & Pelachaud, 2002; Stone, 2004; and Pirsiavash & Ramanan, 2012. These taxonomies are discussed in detail in Section 2.5.4.

In the late 1990s and early 21st century, the research focus was shifted from the theory of human actions to other disciplines, including: task analysis (Swezey et al., 1998); computer-supported cooperative work (Robertson, 1997, 2000); human abilities requirements (Cockayne, 1998; Cockayne & Darken, 2004); and computer and mobile user behaviours (Hostetter & Alibali, 2008). With the introduction of computers, terms such as 'human action theory', the 'taxonomy of human actions' and 'basic actions' have changed to terms such as 'touch screen gestures', 'computer-supported cooperative', and 'online customer behaviour'. This change of terminology might be the reason that less research containing these terminologies has been reported in contemporary literature. Instead, the computer-based terminologies have been used increasingly in recent years.

Cockayne and Darken (2004) stated that “human performance research in VEs is being explored by professionals from myriad fields of research: human factors, behavioural and cognitive psychology, computer science, industrial engineering, and biomechanics, to name a few” (p. 406). The interest in human actions in virtual environments requires a common ground and understanding. Taxonomies can help to create a standard language for virtual environment researchers. The new authentic assessment method, ALAM, introduced in Chapter 4 of this thesis, focuses on learners’ actions and performance mainly in (2D/3D) VTEs. The automated assessment of Action-based learning scenarios requires the processing of learners’ actions. Due to the automation characteristic of ALAM, the use of taxonomy of human actions to standardise the process is vital for consistency of results that ensures the internal reliability of this method (Phelan & Wren, 2006).

In the following sections, research studies and emergent findings are reviewed in order to discover a suitable taxonomy for human actions or the foundation for a new taxonomy as a common ground for studies involved in human actions and performances, especially simulated human actions in (2D/3D) VTEs used by the ALAM framework are developed and introduced in this thesis. However, the author believes that this taxonomy should be open and adaptable so that it can be employed in any field of study and not only in 3D VTEs or the research conducted for this thesis.

2.5.1. Human Action, Activity, Behaviour, Task, and Performance

When reviewing the literature on human actions, one encounters the terms ‘action’, ‘activity’, ‘behaviour’, ‘task’, and ‘performance’ used in relation to “human actions”, but there are differences between these terms.

Turaga et al. (2008) claimed that the terms ‘action’ and ‘activity’ are used interchangeably in the computer vision literature and shed light on the issue by providing a definition for both terms. Actions are “simple motion patterns usually executed by a single person and typically lasting for short durations of time” (p. 1473), while activities are “the complex sequence of actions performed by individuals or achieved by several humans who could be interacting with objects, the environment or each other in a constrained manner. They are typically characterized by much longer temporal durations" (p. 1473). The difference between actions and activities can be exemplified by comparing running or/and shooting a ball with a football team scoring a goal. Although Turaga et al. (2008) concentrated on the use of these two terms in the field of computer vision and action recognition in videos, the issue of the interchangeable use of the two terms exists in other fields of research. Allen (1984) defined action in the field of artificial intelligence as "an occurrence caused in a 'certain' way by the agent" (p. 138). Kuutti (1995) defined activities as "individual and cooperative actions, and chains and networks of such actions, related to each other by the same overall object and motive. Participating in an activity is performing conscious actions which have an immediate, defined goal" (p. 26). In this research, 'action' is an atomic function executed by entities in a specific context. Entities can vary from humans to computerised agents.

Behaviour is the variety of actions and mannerisms executed by people in tandem with themselves or their environment, which includes the other people around them as well as the physical environment. It is the reaction of a person to various stimuli, whether internal or external, conscious or subconscious, explicit or implicit, and intentional or unintentional (Minton and Khale, 2014).

Hogan et al. (1990: p. 2) defined tasks as the "overt and/or covert human actions (process) that begin in response to a set of initiating stimuli external to the operator (stimulus) and end with the production of an identifiable output from the operator".

Comparing the definitions of activity, task, performance, and behaviour, the similarity that might have caused the issue of the interchangeable use of terms emerges. However, a more precise scrutiny of the definitions shows the unique characteristics that differentiate these terms from each other. While activity is a specific set of actions performed with a goal, the behaviour is mainly in response to various stimuli. Like behaviour, tasks are in response to a set of stimuli, although they are different in that behaviour is in response to both internal and external stimuli, while the task is in response to external stimuli.

Although the terms 'action', 'activity', 'behaviour', 'performance', and 'task' are defined differently, they all involve human actions. Hence, the existing taxonomies related to human actions, activities, behaviours, performances, and tasks are studied in Section 2.5.4.

2.5.2. Context and Interpretation

Tuomela (1977) looked at different consequences of the same action in different contexts. He showed the difference between the 'semantical and conceptual relationships' of the same action. As an example, he used agent A flipping the switch which in different cases produces different consequences (Figure 2.6):

- A's hurting his finger;
- A's exploding a bomb;
- A's turning on the light;
- A's illuminating the room;
- A's alerting a prowler.

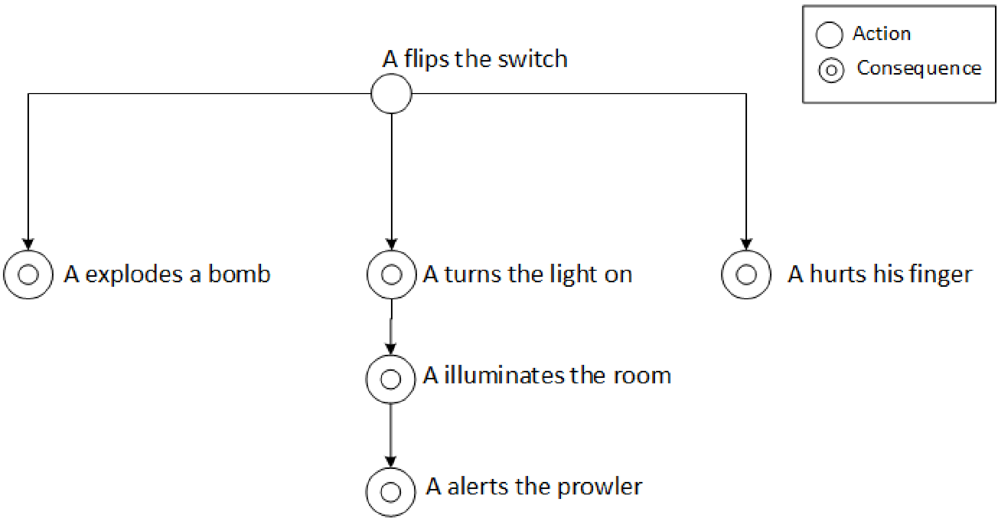


Figure 2.6: Agent A flipping the switch (Tuomela, 1977)

All these consequences are caused by the same basic action of flipping the switch. Actions are 'context-dependent' which makes it very difficult to classify them. Here, Mills (1940) used the term 'Situated Action' based on the language of motive for an action. Suchman (2007) recognised plans as the main influential force behind every purposeful action. She re-introduced the term 'Situated Action', emphasizing "that every course of action depends in essential ways on its material and social circumstances" (p. 70). However, situated action is not considered as a certain type of action but, as Chen & Rada (1996: p. 0) stated, "the term situated actions emphasizes the interrelationship between an action and its context of performance". As consequences might not be the expected outcome, they are classified regardless of the basic action that caused the consequences. The consequences might be classified as an independent individual action, e.g. exploding, or as a complex action constructed by a sequence of actions (Section 5.2.1).

Time is one of the factors to be considered when interpreting human actions, especially when assessing performed actions. The amount of time humans spend wandering around in the environment, or their reaction time to an event, can be interpreted in many different ways. For different reasons, humans show delays in taking action, and these delays may be due to the lack of knowledge, wondering, remembering, or decision-making. If learners are in their first stage of learning, depending on the previous number of attempts, their performance is characterised by slow and irregular actions since they continuously need to use feedback (Holding, 1989). The novice learner performs an action and visually observes the consequences, and then performs another action, assesses the outcomes again, and so on. “With practice, however, the sequencing and timing of [actions] seem to shift from direct visual control to an internal form of control. As a result, performance appears to be rapid and coordinated” (Holding, 1989: p. 49).

2.5.3. Taxonomy and Different Fields of Use

Fleishman et al. (1984: p. 21) define taxonomy as “the theoretical study of systematic classifications including their bases, principles, procedures, and rules” that includes ‘Classification’ and ‘Identification’. ‘Classification’ is the ordering and arrangement of entities into groups or sets on the basis of their relationships (Simpson, 1961), while ‘Identification’ is defined as “the allocation or assignment of additional, unidentified objects to the correct class, once such classes have been established by prior classification” (Fleishman et al., 1984: p. 21).

Fleishman et al. (1984) investigated taxonomies in different human-related areas (Section 2.5.4) including: Clinical classification; Personality classification; Classification systems based on biographical data; Environments and situations; Education; Organizational behaviour; Work motivation; and Team functions. However, these are not the only fields using taxonomies and classifications related to human actions, behaviours, tasks, and performances. After personal computers, and especially the Internet, had become pervasive, a vast number of new tools were developed to study behaviours of consumers. These consumers vary from housewives shopping for groceries online to learners earning a higher degree online. The development of these new tools opened the door to new research fields such as

Human-Computer Interaction (HCI), e-Commerce, website visitors' behaviour, and e-Learning with an interest in human behaviours and their classification (e.g. Robertson, 2000; Lim, 2003; Graf et al., 2009).

If a taxonomy is to be used in different fields of study, it should be flexible, transferable, and exhaustive. It should accommodate changes and extensions; be adaptable to different fields instead of focusing on a certain problem at hand; and finally, be exhaustive by covering different actions from different aspects of human life and not only a fraction of it.

2.5.4. Review of Current Taxonomies

2.5.4.1. Actions

Goldman (1970) explained the classification of human actions in his theory of human actions. His discussion of individuation, Act-type, Act-token, basic and non-basic actions informed his explanation of the structure of actions and level-generation. His theory of human actions included his views on intentional actions and wants. He criticised the way that 'identity thesis[1]' has been used by other researchers such as Anscombe (1958), Donald Davidson (1967), Shwayder (1965), and D'Arcy (1963) in the 'fifties and 'sixties, and how they assume two different acts like 'moving right arm' and 'move queen to king-knight-seven' are the same or recognise both as basic actions based on 'same identity' thesis. He states: "moving my hand is a basic action, whereas checkmating my opponent and turning on the light are not basic actions. Rather, they are actions I perform by performing some basic actions" (Goldman, 1970: p. 6).

Tuomela (1977) did not provide a precise classification of actions, although he suggested and defined different types of actions such as basic action, bodily action, generated action, token action and complex actions. Bodily actions are body movements causing basic actions. Action token is defined as "(possibly complex) singular events", which "has the structure of a finite sequence of events <v,...,b,...,r>" (Tuomela, 1977: p. 290). Tuomela acknowledged Goldman's theory of human actions as being the most detailed, but nevertheless believes that "his definition of a basic

[1] Goldman calls the Davidson's thesis (Goldman, 1970: P. 2, Lines 9-11) identity thesis as he recognises the identity to be the relation between the two acts that he believes are identical, for example, pointing the gun and pulling the trigger, and shooting the victim. For more information about identity thesis please refer to Cacioppo & Tassinary (1990: pp. 19-20).

action is faulty, and he does not, after all, give any detailed account of the central problem of how complex actions are supposed to be built out of (simple or compound) basic or bodily actions" (Tuomela, 1977: p. 291). Tuomela contended that Goldman's theory of human action was faulty because just wanting something will not result in any action at all.

2.5.4.2. Activities

To develop a driving training device implications system, Willis (1961)[1] developed a three-level hierarchical classification scheme for task behaviour. The first and highest level of the activities in Willis' scheme consists of: a) Receptor Activity (input), b) CNS Activity (black-box), and c) Effector Activity (output). The input activities are divided into verbal and non-verbal cues, each of which can be in the form of detection or identification/recognition. This is while the black-box activities recall (data, facts, and procedures) or manipulate and process the data (decision making). The output activities, including skilled motor acts and overt verbalisation, are classified as different movements, reactions, and different types of verbalisations (shown in Appendix 2).

Kuutti (1995) investigated the Activity Theory[2] as a potential framework for human-computer interaction. He argued that as activities are dynamic, the border between activity and action is a blur. He also emphasised this dynamic character with an example: "a software project may be an activity for the team members, but the executive manager of the software company may see each of the projects as actions

[1] Due to the issue of digitising old documents, most of the early taxonomies and classifications were not available during the research conducted for this thesis. As a result, the author had to use the book by Fleishman et al. (1984) that can be considered as a thorough review of prior research on taxonomies of human actions and performances. The cited research studies include: Willis (1961), Berliner et al. (1964), Reed (1967), Oller (1968), Finley et al. (1970), and Bennett (1971). Of these taxonomies, the only available publication is that of Bennett (1971). These taxonomies are introduced in the following paragraphs.

[2] "Activity Theory has long historical roots which are quite unfamiliar to most Anglo-American readers. The oldest background tradition — the 18th and 19th century classical German philosophy from Kant to Hegel — has remained distant because that tradition opposed the emerging (British) empiricism that was later to become the foundation of mainstream Anglo-American scientific thought. The classical German philosophy emphasised both developmental and historical ideas and the active and constructive role of humans. Another root – also alien to many – consists of the writings of Marx and Engels, who elaborated on the concept of activity, and the third source is the Soviet cultural-historical psychology, founded by Vygotski, Leontjev and Lurija. Activity Theory was first born within Soviet psychology, but today there is an emerging multidisciplinary and international community of scientific thought united by the central category of activity — a community reaching far beyond the original background. Broadly defined, Activity Theory is a philosophical and cross-disciplinary framework for studying different forms of human practices as development processes, both individual and social levels interlinked at the same time" (Kuttie, 1995: pp. 22-33)

within his or her real activity at the level of the firm" (p. 27). His hierarchical levels of activity are shown in Figure 2.7:

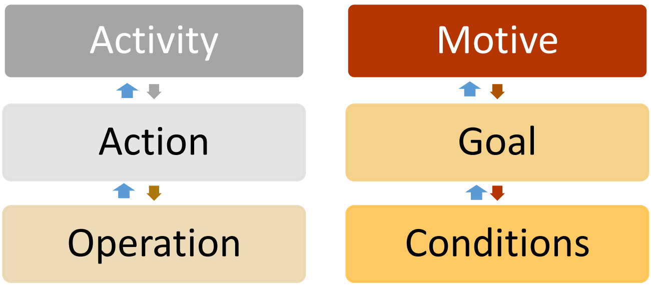


Figure 2.7: Hierarchical levels of activity (Kuutti, 1995)

Aggarwal & Ryoo (2011) conceptually categorised human activities according to four different levels, based on their complexity: Gestures, Actions, Interactions, and Group Activities. They saw 'gestures' as "elementary movements of a person's body part" that is "the atomic components describing the meaningful motion of a person" (e.g. stretching an arm, raising a leg) (p. 16:2). A set of temporally organised gestures performed by an individual is called an 'Action' (e.g. walking, waving, and punching). Aggarwal & Ryoo (2011) introduced 'Interactions' as "human activities that involve two or more persons and/or objects" (p. 16:2). For example, an interaction can be between two humans, such as two persons fighting, or it can be a human-object interaction involving two humans and one object, such as a person stealing a suitcase from another. The last level in the Aggarwal & Ryoo (2011) classification comprises group activities. These are "the activities performed by conceptual groups composed of multiple persons and/or objects" (e.g. a group of persons marching, a group having a meeting, and two groups fighting) (p. 16:2).

Pirsiavash and Ramanan (2012) proposed a very subject-specific taxonomy for daily living activities (Figure 2.8) from the first-person camera view. This taxonomy classified the activities of daily living (ADL) into three main classes: Hygiene, Food, and Entertainment.

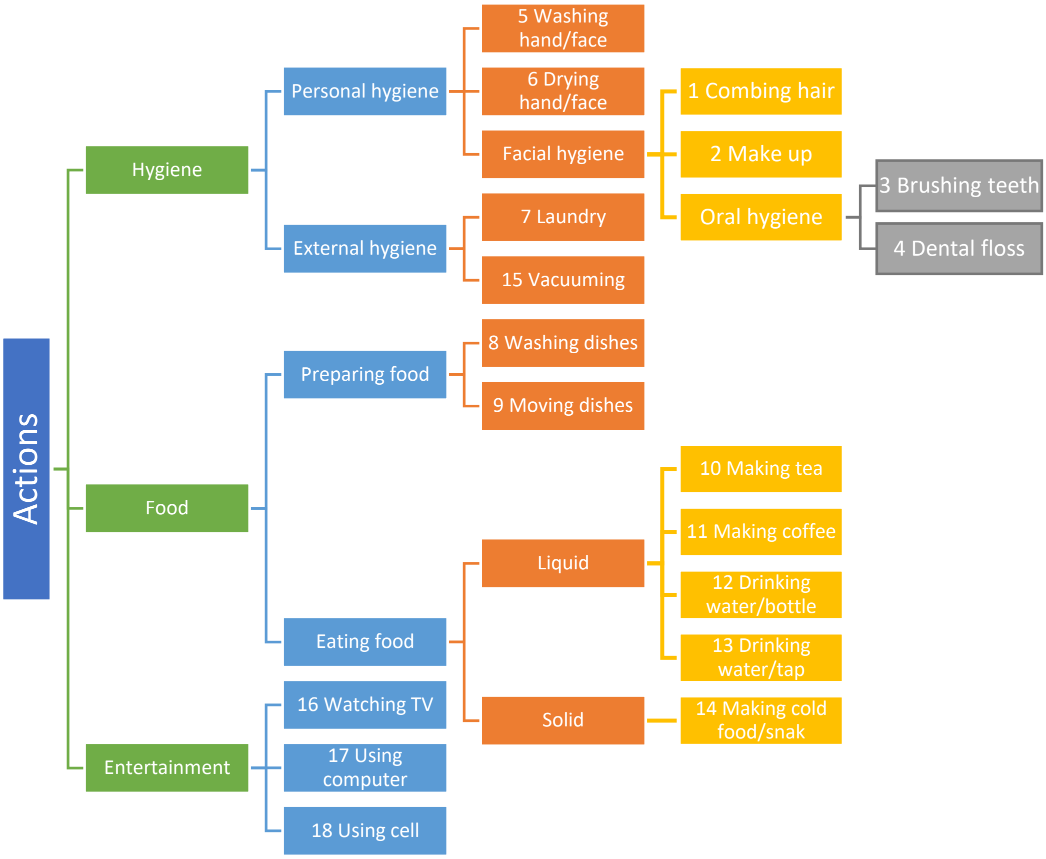


Figure 2.8: Manually-designed functional ADL taxonomy (Pirsiavash & Ramanan, 2012)

2.5.4.3. Gestures

Gestures as actions have been categorised and classified by researchers in various fields such as HCI, mobile communication and video lectures and conferences. Karam and Schraefel (2005) performed an extensive literature review of 40 years of research and categorised gestures according to: 1) gesture styles; 2) categorisation by gesture enabling technology; 3) categorisation by application domain; and 4) categorisation by system response, as shown in Appendix 13.

Action recognition in videos has been well-developed over recent decades, and recognition of human gestures is one of its focuses. Educators use many gestures in teaching. Zhang et al. (2010) found 866 gestures and identified 126 fine equivalence classes which were further clustered into nine semantic classes. They also annotated all the found and classified gestures. These nine clusters are:

1. Put
2. Spread
3. Swipe
4. Close & Open

5. Flip & Swing
6. Touch
7. Pointing
8. Hold
9. Others

Although Zhang et al. (2010) developed their taxonomy of gestures in teaching and an annotation system for it, nevertheless their work benefited from McNeill (1992)'s classification of gestures in language and communication which itself was based on the very early works of other researchers shown in Table 2.4.

Table 2.4: Four Gesture Classification Schemes (McNeill, 1992)

Present Categories	**Efron (1941)**	**Freedman and Hoffman (1967)**	**Ekman and Friesen (1969)**
Iconics	Physiographics Kinetographics	Literal-reproductive	Kinetographs Pictographs
Metaphorics	Ideographics	Concretization Minor and major qualifying	Ideographs Underliners Spatials
Deictics	Deictics		Deictics
Beats	Batons	Punctuating	Batons Rhythmics
Butterworths		Speech Failures	

As with human behaviour classifications, the taxonomy of gestures was inclined towards their technological use. Human gestures in relation to touch screens, smartphones, and surface computers have been the new research focus in recent years. Epps et al. (2006), Wobbrock et al. (2009), Freeman et al. (2009), Urakami (2012), and Piumsomboon et al. (2013) provided some of these classifications.

2.5.4.4. Virtual and Computerised Environments

Computer-supported Cooperative Work (CSCW) is one of the disciplines using collaborative virtual environments and following that, Embodied Actions. Robertson (1997, 2000) created a taxonomy of Embodied Actions for cooperative design in a distributed company. Like other researchers, he developed this taxonomy according to his research needs, stating: "the taxonomy presented in this paper was developed as a possible bridging structure between the field study of cooperative work in practice and the design of technology that might support that work over distance" (p. 208). Robertson specified the categories in an open and flexible manner so that people could adopt this taxonomy in practice. The taxonomy divides the embodied actions into Individual Embodied Actions, and Group Activities constituted by individual embodied actions. Individual embodied actions are in relation to physical objects, other bodies, and the physical workspace. In relation to physical objects, Robertson classified Embodied Actions into moving physical objects, producing a private physical representation, highlighting some aspect of an object, and personal use of a physical object. In relation to other bodies, the classification included emitting signs and monitoring signs, and pretending to be another body. Finally, in relation to the physical workspace, moving around, pointing at something, shifting the direction of gaze, and moving in or out of the shared space are different classes of embodied actions. Group activities constituted by individual embodied actions are shaped by conversing, looking at the same thing at the same time, organizing shared communication resources, creating a shared representation, shared physical use of an object, focusing group attention, breaking into smaller groups and reforming, seizing the moment, and doing something else.

Cappella & Pelachaud (2002) studied human behaviour to build virtual interactions based on their human counterparts. Their archive included 100 interactions, including “same-sex and opposite-sex pairs, dyads with longer histories (greater than six months as friends) and strangers, partners with similar and different attitudes, and expressive and reticent pairs”. They coded the interactions and used these in the system to analyse behaviour. They studied and coded interactions that included vocalic behaviours, eye gaze, smiles and laughter, head nods, back channels, posture, illustrator gestures, and adaptor gestures.

The Rapid Assessment of Tasks and Context (RATaC) taxonomy (Stone, 2004) (shown in Appendix 8) was developed to reduce the restrictions on logistics, timing and finance that human factors experts face when trying to find the essential constituents of training scenarios in order to define the scope of technology-based training solutions. Although the RATaC taxonomy was used successfully in case studies such as the Tornado F3 Avionics Training Facility (ATF), Naval Gunnery and Helicopter Voice Marshalling semi-immersive VR trainers, keyhole surgery and temporal bone procedure, TBT systems and a submarine qualification training and submarine rescue TBT system, Stone (2004) acknowledges that this taxonomy is somewhat limited, and its development does not seem to have been informed by established scientific theory.

Cockayne (1998) and Cockayne & Darken (2004) based their taxonomy for two-handed, whole-hand input and locomotion in virtual environments, on the Fleishman Job Analysis Survey (F-JAS) (Fleishman and Reilly, 1992). They created a flowchart showing how to develop a classification of Human Abilities Requirements (HAR) (shown in Appendix 9).

Bloomfield et al. (2003) introduced the Haptic Action Taxonomy (shown in Appendix 12) which was meant to cover only manual actions, not all possible human actions, simulated in virtual environments by haptics. These actions do not cover the full body actions as noted by the authors who stated (Bloomfield et al., 2003: p. 226):

> "these actions are mostly arm and hand actions, as that is what the majority of current haptic research, as well as current available haptic devices, focuses on. The classification consists of actions requiring fine motor control, significant arm strength, tactile friction, cooperative two-handed tasks, braced two-handed tasks, manipulating a deformable object, tool-assisted tasks, and multiple finger tasks".

Verhulsdonck (2009) opined that a common taxonomy for non-verbal behaviours in VW needs to "include both rhetorical acts (actions of choice), as well as those that are procedurally driven by the utterances or the psychological state of the avatar" (p. 8). He suggested that the developed standards should openly allow such advancement. Hostetter and Alibali (2008) defined gestures as simulated actions. They created a framework based on that definition, which "asserts that gestures emerge from the perceptual and motor simulations that underlie embodied language and mental imagery" (p. 502).

Hurst (2011) classified recorded data into (1) reflective data, (2) machinima and (3) virtual environment data. Hurst identified the last category as the most relevant, covering the recording of the avatars' in-world actions and their interactions with objects. However, her paper was not published, and the research was abandoned (cited in Chodos et al., 2014).

Chodos et al. (2014) created a definition of an Avatar Capabilities Model (ACM) and categorised students' avatar actions according to three pedagogically based themes: movement, experiencing the world, and social interaction. The ACM model is shown in Table 2.5.

Table 2.5: Definition of Avatar Capabilities Model (by permission from Chodos et al., 2014)

Action = < Movement | Sensing | Object Manipulation | Communication >
Movement = < Move | Sit >
Move = < Actor, Movement Type, Start Location, End Location >
Sit/Stand = < Actor, Sit/Stand Type, Sit/Stand Location>
Sensing = < Actor, Modality, Target >
Object Manipulation = < Create | Hold | Transfer | Take | Interact >
Create = < Actor, Created Entity >
Hold = < Actor, Held Entity >
Transfer = < Actor, Target, Transferred Entity >
Take = < Actor, Taken Entity >
Interact = < Actor, Entity, Message, Options, Choice, Response >
Communication = < Speak | Write | Gesture >
Speak = < Actor, Message >
Write = < Actor, Message >
Gesture = < Actor, Communication Type, Description >

2.5.5. Task, Skill, Behaviour, and Performance

Berliner et al. (1964) used a three-level classification (shown in Appendix 3) for selecting an optimal method for performance measurement in military jobs. Their classification included 'Processes' (the highest level), 'Activities', and 47 'Specific behaviours'. These 47 specific behaviours "were selected for being (1) reliably identifiable, (2) simple acts with quantifiable properties, and (3) involved in a variety of jobs" (Fleishman et al., 1984: p. 93). Berliner et al. (1964) did not assess the reliability of their classification; hence, this precludes the real-world use of this classification. Similarly, the same classifications were developed by Reed (1967) and Oller (1968) when compiling "a glossary of 130 action verbs, each defined as precisely as possible, along with a cross-referenced system of synonyms; a similar glossary, minus the synonyms, is provided for nouns" (Fleishman et al., 1984: p. 96).

The next classification mentioned by Fleishman et al. (1984) is the Meister Taxonomy (Shown in Appendix 4). The Meister Taxonomy was developed by Finley et al. (1970) to be used in the "man-machine" system of a hypothetical multi-crew, Extended Earth Orbit Scientific Laboratory. "This approach actually includes four classificatory systems, three representing different levels of description of human behaviour and the fourth representing dimensions of task characteristics" (Fleishman et al., 1984: p. 101).

The other relevant taxonomy studied by Fleishman et al. (1984) was Bennett’s Semantic Classificatory Approach (1971). In their pretesting stage, they chose 20 tasks to be judged in order to find the measure of the familiarity and also, the relevance to three hypothesised dimensions: ideas, people, and things. Subsequently, ten tasks were selected. Also, 25 out of 200 verbs were ranked by 36 students so as to measure the applicability of these 25 verbs to the ten pretested tasks using a 4-point scale ranging from “not at all applicable” to “extremely applicable” (Bennett, 1971: p. 231). The tasks are investigated and ranked according to four factors: Cognitive, Social, Procedural, and Physical. The results of the final taxonomy are shown in Appendix 5.

Fleishman (1975) identified six categories of tasks in the literature: identification, discrimination, sequence learning, motor skill, scanning, and problem-solving. He contended that it is important to be clear about why we are interested in task classification because those who create these classifications do not view them as an end, but more as a tool in their research to improve their capability to interpret or/and foresee different aspects of human performance. He stated: "We can elect to develop a system of classification having utility for a limited area (e.g., the classification of tasks with respect to which particular training methods are found most effective in promoting high levels of task performance), or we may look for a system from which a variety of applications may stem" (Fleishman, 1975: p.1128).

Fleishman (1975, 1982) examined various taxonomies and classifications studies and identified four bases for classifying tasks: "behavior description approaches (e.g., handling objects, analyzing data), behavior requirements approaches (e.g., problem solving, scanning), ability requirements approaches (e.g., spatial-visualization ability), and task characteristics approaches (e.g., type of display, instructions, goals)" (1982: p. 827). He also devised different procedures for internal and external validity evaluation of task classification systems. The Manual for the Ability Requirement Scales (MARS) is one of the outputs of their research. Fleishman and Reilly (1992) developed a job analysis survey based on 52 human abilities studied in their earlier research (shown in Appendix 6).

Swezey et al. (1998) developed a methodology named Task and Training Requirements Analysis Methodology (TTRAM). When developing this methodology, they needed to create a taxonomy of tasks and skills. Their behavioural classification system (Shown in Appendix 7) was based upon the work of Berliner et al. (1964) and

Cannon-Bowers et al. (1995) and was used to “categorize and decompose performance requirements associated with each sub-task. Skill components underlying each sub-task are classified into the skill/knowledge categories” (p. 1689). The taxonomy includes both individual and team-oriented skill classification categories. These categories include decision-making, communication, leadership, and situational awareness as suggested by Cannon-Bowers et al. (1995).

Salmon et al. (2008), in reviewing the RATaC, stated that “in order for the method to be both valid and comprehensive, much more detailed taxonomy of tasks and task contexts is required” (Salmon et al., 2008: p. 6). Salmon et al. (2008) considered the context of human actions crucial and reviewed the literature extensively for “Performance Shaping Factor Taxonomies”, which is shown in Appendix 10. Also introduced by Salmon et al. (2008) is the taxonomy of tasks and taxonomy, and the Performance Shaping Factor (PSF), as shown in Appendix 11.

2.5.6. Discussion

A summary of the reviewed literature distinguishes the following main points:

Researchers tend to focus more on developing taxonomies of human behaviour, activity, performance, skill, and task (sections 2.5.4 – 2.5.5), rather than the human actions (Section 2.5.4.1), both computerised (online/offline) or non-computerised. It seems that the application of those taxonomies in the industry has contributed to this matter.

Embodiment and virtualisation do not seem to attract much research in computerised human action classification and recognition, while action recognition is mainly associated with computer vision research (e.g. subtitle synchronisation, surveillance systems, and patient monitoring systems) in the literature. However, fields such as CSCW and HCI use classification of human actions in virtual environments in a very focused and limited way, and there is still no exhaustive taxonomy in these fields (Section 2.5.4.4).

The majority of the taxonomies focus on a specific (research) problem, and there is no intention to develop further extensions and adaptations.

All the actions that have been considered are mostly in relation to other humans, their environment, and objects.

With regards to the points mentioned above, one can conclude that there is a need for a taxonomy that is flexible, transferable, and exhaustive. For a taxonomy to be flexible, it should be able to evolve in response to the introduction of new applications or technologies. A taxonomy should allow adaptation and transfer to other scenarios and contexts, so different fields of research can benefit from it. The lack of a taxonomy that is not constrained by the original application and is exhaustive enough to cover a diverse range of applications is observed in the literature. Table 2.6 summarises the taxonomies investigated in this section.

Table 2.6: Published taxonomies of human actions, performances, tasks, and behaviour

	Taxonomy	Field	Literature	Main characteristics	Missing part
1	Theory of human actions	Human Actions	Goldman, 1970	Act-type, Act-token, basic and non-basic actions	Too general
2	No name	Human Actions	Tuomela, 1977	Basic action, bodily action, generated action, token action and complex actions	No precise classification of actions
3	Driving training device implications system	Task and behaviour	Willis, 1961	Receptor Activity (input), CNS Activity (black-box), Effector Activity (output)	No classification of actions
4	No name	Performance	Berliner et al., 1964	Processes (the highest level), Activities, and 47 Specific behaviours	No classification of actions
5	No name	Human behaviour	Reed, 1967 & Oller, 1968	A glossary of action verbs and nouns	No classification of actions
6	Meister Taxonomy	Human behaviour and task analysis	Finley et al., 1970	Four classificatory systems: three representing different levels of description of human behaviour and the fourth representing dimensions of task characteristics	No classification of actions
7	Semantic Classificatory Approach	Tasks	Bennett, 1971	10 tasks and 25 verbs ranked considering four factors: Cognitive, Social, Procedural, and Physical	No classification of actions
8	Task and Training Requirements Analysis Methodology (TTRAM)	Tasks and skills	Swezey et al., 1998	Individual and team-oriented skill classification categories: decision-making, communication, leadership, and situational awareness	No classification of actions
9	Performance Shaping Factor (PSF)	Tasks and performances	Salmon et al., 2008	Organisational, Environmental, Task, Personal, Workspace, Temporal, Social	No classification of actions
10	The Manual for the Ability Requirement Scales (MARS)	Skills	Fleishman & Reilly, 1992	A job analysis survey based on 52 human abilities studied earlier in their research	No classification of actions
11	No name	Activity	Kuutti, 1995	Activity, Action, operation	No classification of actions
12	No name	Activity	Aggarwal & Ryoo, 2011	Gestures, Actions, Interactions, and Group Activities	No classification of actions
13	No name	Gestures	Karam & Schraefel, 2005	Categorisation by: gesture styles, gesture enabling technology, application domain, and system response	Limited to gestures
14	Classification of gestures in language and communication	Gestures in language and communication	McNeill, 1992	Iconic, metaphoric, deictic, and beat gestures	Limited to gestures

	Taxonomy	Field	Literature	Main characteristics	Missing part
15	Taxonomy of gestures in teaching	Gestures in teaching	Zhang et al., 2010	Put, spread, swipe, close & open, flip & swing, touch, pointing, hold, and others	Limited to gestures
16	Taxonomy of Embodied Actions for cooperative design in a distributed company	Computer-supported cooperative work (CSCW)	Robertson, 1997, 2000	Individual Embodied Actions: in relation to physical objects, other bodies, and the physical workspace. Group activities: conversing, looking at the same thing at the same time, organizing shared communication resources, creating a shared representation, shared physical use of an object, focusing group attention, breaking into smaller groups and reforming, seizing the moment, and doing something else.	Designed from the interaction perspective more than action. The (inter)actions are defined "in relation to[1]" something in a general way.
17	No name	Virtual interactions	Cappella & Pelachaud, 2002	Vocalic behaviours, eye gaze, smiles and laughter, head nods, back channels, posture, illustrator gestures, and adaptor gestures	Limited to gestures
18	Classification of Human Abilities Requirements (HAR)	Hand input and Locomotion in VW	Cockayne & Darken, 1998, 2004	Two-handed whole-hand input: gross touch, fine touch, force reflection, temperature discrimination, pain discrimination, control precision, multi-limb coordination, arm-hand steadiness, manual dexterity, finger dexterity, wrist-finger speed, and speed of limb movement Locomotion: walk, jog, acceleration from rest to a walk or jog, deceleration to rest from a walk or jog, acceleration from walk to jog, deceleration to walk from jog, turning in place (no forward or lateral movement), sidestepping (purely lateral movement), tilting upper body without foot movement	Limited to gestures and locomotion on omni-directional treadmill
19	Activities of Daily Living (ADL)	Action recognition in videos	Pirsiavash & Ramanan, 2012	Daily actions: Hygiene, Food, and Entertainment	Limited to very simple and limited number of daily actions
20	No name	Avatar actions in VW	Hurst, 2011	Reflective data, machinima, virtual environment data	No information available

[1] "The 'in relation to' recognises the indexicality of all embodied actions. Indexicality, in this context, is not used in a narrow linguistic sense, but in the ethnomethodological sense that all actions need to be interpreted within the context in which they occur" (Robertson, 1997: p. 211)

	Taxonomy	Field	Literature	Main characteristics	Missing part
21	Haptic Action Taxonomy	Manual actions simulated in VE by haptics	Bloomfield et al., 2003	Manual actions (related to arm and hand) requiring: fine motor control, significant arm strength, tactile friction, cooperative two-handed tasks, braced two-handed tasks, manipulating a deformable object, tool-assisted tasks, and multiple finger tasks.	Limited to manual actions related to arm and hand to be simulated in a VE with haptics
22	Rapid Assessment Of Tasks and Context (RATaC)	Virtual Training	Stone, 2004	Human operator involvement/role, interaction style, task/workplace physical coupling, technology appropriateness, content, fidelity	No classification of actions
23	Avatar Capabilities Model (ACM)	Learning in VW	Chodos et al. , 2014	Actions: movement, object sensing and manipulation, communication	Sensing and manipulation are only towards objects. Actions that enable the actor to operate or make decisions are not considered.

2.5.7. Evaluation of a Taxonomy

As with any other research output, or designed product, a newly-developed taxonomy needs to be evaluated for validity. A taxonomy has to be valid both internally and externally for it to be used with confidence, given that if the taxonomy as a tool is not reliable and valid, the results are not usable. Although there is an adequate amount of literature on the evaluation of website taxonomies and ontologies, there is a paucity of available information on the evaluation of human actions taxonomies. The literature on the validation of taxonomies is mostly from the ’80s even though it is still regularly cited and used in various types of research.

Fleishman et al. (1984) and Fleishman and Mumford (1991) describe three primary criteria for taxonomy evaluation including internal validity, external validity, and use rate. If the classification “is logical and parsimonious within itself”, and “is capable of accomplishing its intended purpose”, it is identified as a valid classification (Fleishman et al., 1984: p. 82). The use rate of the taxonomy by scientists is not so easy to determine as the taxonomy should be published and used first. However, high internal and external validity for high-quality ‘human engineering’[1] increases the use rate among scientists. As for the internal validity, two criteria concern us the most: having “mutually exclusive classes on the horizontal level” which places each entity under just one class, and the second is its being “exhaustive” which enables every entity to fall under a class one way or another (Fleishman et al., 1984). The mutually exclusive classes criterion is mostly satisfied in “monothetic quantitative systems”[2] (Fleishman et al., 1984: p. 83), known as monothetic classes (Bailey, 1994), and is the hardest to satisfy in the qualitative systems such as behavioural taxonomies, performance taxonomies and the taxonomy of human actions. The external validity of the taxonomy is used to evaluate its generalisability and transferability, which enables other researchers to use the taxonomy in their research with or without alteration. Different researchers have used various methods to validate and evaluate their taxonomies. Some of this research is described in the following paragraphs.

[1] Science dealing with the application of information on physical and psychological characteristics to the design of devices and systems for human use. This term is an alternative for ‘ergonomics’ (Holstein, n.d.).

[2] Bailey (1994) introduced 'Monothetic classes' as: "classes containing cases that are all identical on all variables or dimensions being measured. Typologies generally contain only monothetic classes (i.e., a type is a monothetic class)" (p. 7).

Fleishman introduced his taxonomy of human performance in 1975 and continued to develop it during the following years (Fleishman, 1975, 1982). Fleishman et al. (1984) discussed additional validity testing methods for the taxonomy of human task performance. Fleishman and Mumford (1991) suggested criteria for addressing inferential issues in the evaluation of construct validity of systems assessing the requirements of human task performance. For this, they evaluated Fleishman's (1975, 1982) ability requirement taxonomy and its associated job analysis system, the Manual for the Ability Requirement Scales (MARS) using these criteria. Grobe and Hughes (1993) used prototyping to investigate the validity of their taxonomy in terms of 'validity as value', 'validity as correspondence' and 'validity as robustness'. They examined the substantive, conceptual and methodological aspects of their study. Stone (2004) did not reflect on taxonomy evaluation, but provided several case studies in which RATaC taxonomy was applied, thus showing the applicability of his taxonomy. Lester et al. (2005) used experimentation and data analysis to show that their approach is valid. To compare retention, learning speed and preference of learning with ShadowGuides, a system for learning multi-touch and whole-hand gestures on interactive surfaces was used, against the control condition of learning with video instructions. Freeman et al. (2009) conducted a between-subjects experiment as an evaluation method. Zhang et al. (2010) used expert opinion to evaluate their taxonomy of gestures by asking three experts to use the taxonomy and provide feedback. Urakami (2012) used experimentation as an evaluation method for his taxonomy. A quasi-experimental design was used with the two groups, non-technical and technical. T-tests were conducted for independent samples (comparing experts and novices) and dependent samples (comparing hand shape and motion path). Pirsiavash and Ramanan (2012) used mathematical proof evaluation method including leave-one-out cross-validation and average precision to evaluate object detection accuracy of their taxonomy. Mokkink et al. (2010) and Michie et al. (2013) both used Delphi methods including survey and feedback to create valid taxonomies. Michie et al. (2013) used two rounds of surveys and feedback to show the validity of the Behaviour Change Technique Taxonomy, while Mokkink et al. (2010) used four rounds of survey to measure the degree of agreement among the experts on terminology and definitions of measurement properties in a taxonomy of measurement properties for the evaluation of health instruments.

Table 2.7 summarises various evaluation methods presented in the literature on taxonomies of human actions.

In Chapter 3, various evaluation methods for artefact evaluation in Design Science Research by Vaishnavi and Keuchler (2009) are presented. These include Demonstration, Experimentation, Simulation, Using Metrics, Benchmarking, Logical Reasoning, and Mathematical Proofs. Considering these evaluation methods, and evaluation requirements for taxonomies by Fleishman et al. (1984) and Fleishman and Mumford (1991), and comparing them with different evaluation methods used in the literature of taxonomies (Section 2.5.4), it can be concluded that these evaluation methods are still valid and widely used in different research.

Table 2.7: Taxonomy validating methods in different research

Research	Taxonomy	Validity tests
Fleishman et al. (1975, 1982, 1991)	Manual for the Ability Requirement Scales (MARS)	Experimentation Case study Using Metrics
Grobe and Hughes (1993)	Taxonomy of Nursing Interventions	Prototyping
Stone (2004)	RATaC Taxonomy	Case study
Lester et al. (2005)	Modelling Human Activities	Experimentation Mathematical Proofs
Freeman et al. (2009)	Taxonomy for the Space of Whole-hand and Multi-touch Gestures	Between-subjects Experiment
Zhang et al. (2010)	Taxonomy of Gestures	Expert Opinion
Mokkink et al. (2010)	Taxonomy of Measurement Properties	Delphi Methods – Expert Opinion
Pirsiavash & Ramanan (2012)	Functional Activities of Daily Living Taxonomy	Mathematical Proof
Urakami (2012)	Human-Based Gesture Vocabulary	Quasi-experimental
Michie et al. (2013)	Behaviour Change Technique Taxonomy	Delphi Methods – Expert Opinion

2.6. Summary and Implications

As Sadler (1989) stated, learning assessment is about grading and evaluating the learner's learning outcome, which can be either perceptible (e.g., a report or artwork) or intangible (e.g., skills or knowledge). Research is required to move to advanced assessment methods as the traditional ones (e.g., multiple-choice, closed-answer questions and essays) are too restricted to cope with the flexibility, complexity, and creativity associated with action-based learning (Naidu, 2010). While a range of assessment types may be appropriate as the learner gains mastery, action-based learning implies a level of advancement and ability to execute appropriate actions. Therefore, there must be additional flexibility to cope with the assessment patterns that reflect the actions of learners, particularly at more advanced levels of learning (Wood & Reiners, 2013).

In 3D VTE (e.g., the nDiVE-project as outlined by Reiners et al., 2013), the actions and abilities of learners are primarily 'demonstrated' and must be assessed so that feedback can be provided to the learner. Thus, formative feedback is an important component of the learning process. Using simulations for training in industries is a standard training method these days. Most industries are benefiting from VTEs in numerous ways including cost-effectiveness, safety, and availability. Although new technologies allow the analysis of learners' performance or ongoing feedback during the performance, an automated assessment method that can provide a detailed formative feedback based on performed actions can provide the needed flexibility and lead to more advanced levels of learning. The Action-based Learning Assessment Methodology, discussed and introduced in Chapter 4 of this thesis, provides a framework for achieving this desired flexibility and advancement. This assessment method relies on recognising, codifying, and processing learners' performed actions.

Although action-based learning might not particularly have the answer to all training needs, it can certainly be used as the foundation for a practical learning assessment method focusing on learning from one's mistakes. It also requires learners to perform actions which can reflect their acquired knowledge and the ability to use that knowledge to solve real-life problems. A learning assessment method that relies on human actions needs to be able to recognise and process those actions.

As illustrated in Section 2.5, starting from very first theories of human actions, different researchers have tried to classify human actions, or at least differentiate and define various types of human actions. The initial study of the literature on human actions suggests that the terminology and applications pertaining to these classifications are divided into two eras: before and after computers/networks. Different classifications and taxonomies of human actions were developed in various fields such as activities, tasks, performance, and skills analysis. With the widespread use of computers and computer networks including the internet, the application of human action taxonomies has changed and, consequently, so too has the terminology. Taxonomies of gestures, user website browsing, touch screen gestures, video recognition, computer-supported cooperative work and action learning replaced the human actions theories and taxonomies used in other disciplines. Regardless of these changes, the study of human action taxonomies shows that most of the developed taxonomies are specifically focused, and there is no intention to develop further extensions and adaptations.

The crucial aspect of developing a taxonomy is that it should be evaluated for internal and external validity. The literature on the validation of taxonomies is very limited, and most of the taxonomies are considered valid as they were used in the project for which they were developed and satisfied the needs of the particular research. Fleishman et al. (1984) and Fleishman and Mumford (1991) are found to be the most useful literature on the taxonomy evaluation as these publications are the result of a long period of investigation of taxonomies and classifications of human actions in different fields of research. A review of the literature on the human actions taxonomies showed that those researchers who evaluated their taxonomies also benefited from Fleishman et al. (1984) and Fleishman and Mumford (1991). In this thesis, the researcher did not only rely on these two works, but investigated the literature of human actions taxonomies and compared the evaluation methods used by previous researchers (Table 2.7).

This specified focus in an application leads to a variety of individual taxonomies and classifications that cannot be validated and used for communication between different fields of research. This lack of a standard taxonomy prevents different technologies from communicating easily or at all. There is a need for a standard common taxonomy that is open and flexible, and allows extension, adaptation, and transfer to other scenarios and contexts, and which is not constrained to the original application. Chapter 5 discusses a new taxonomy of human actions namely BEHAVE, that is developed to be flexible, transferable, and exhaustive.

This chapter investigated various action-based learning methods and theories, learning assessment, feedback, authentic and automated assessment. It then investigated and discussed different taxonomies and classifications involving human actions in various fields of research. The next chapter, Chapter 3, introduces the research structure of this thesis, and the research methodology.

Chapter 3: Research Methodology

3.1. Introduction

As mentioned in the Introduction chapter, this research started with the main question asking how learners' goal-oriented actions and action-sequences can be represented, analysed, and automatically assessed in virtual worlds. To answer this question, the initial aim of the research was to examine a method to assess what students have learned by expressing their goal-oriented actions in a simulation through the use of a virtual world environment. The objectives of the proposed research were: to develop an Action-based Learning Assessment System in virtual worlds through representing the goal-oriented actions and action-sequences in virtual worlds; analyse actions and their sequences; assess actions and their sequences; generate an automated formative feedback, and evaluate the efficacy of the generated formative feedback. After the development of ALAM (Chapter 4), the researcher recognised the crucial need for a formalised representation of the learner's actions in order to create consistency in ALAM. Consequently, the focus of the research took a new direction: the development of an exhaustive standard taxonomy of human actions. The change of aim and objectives are demonstrated in Figure 3.1.

This chapter discusses the research questions, research aim, objectives, and limitations that affect the scope of the research. The research methodology and design are also discussed along with the rationale for choosing the research methods and research design. The chapter concludes with a discussion of the participants, the sampling methods, and research ethics.

3.2. Research Question

How can learners' actions be formally represented to create consistency in the assessment process leading to an automated post-performance formative feedback?

3.3. Research Aim and Objectives

3.3.1. Research Aim

This research explores how learners' performed actions can be formally represented to create consistency by comparing learners' actions with expert reference solutions, in order to generate an automated post-performance formative feedback.

3.3.2. Research Objectives

1. To analyse the literature of taxonomies and classification of human actions in different disciplines (Section 2.5).
2. To develop a classification of human actions (Section 5.2.1).
3. To develop a set of action attributes to describe the actions (Section 5.2.2).
4. To design a formal syntax to structure the actions as computer-readable data (Section 5.2.2).
5. To evaluate the taxonomy for internal and external validity (Chapters 6 and 7).

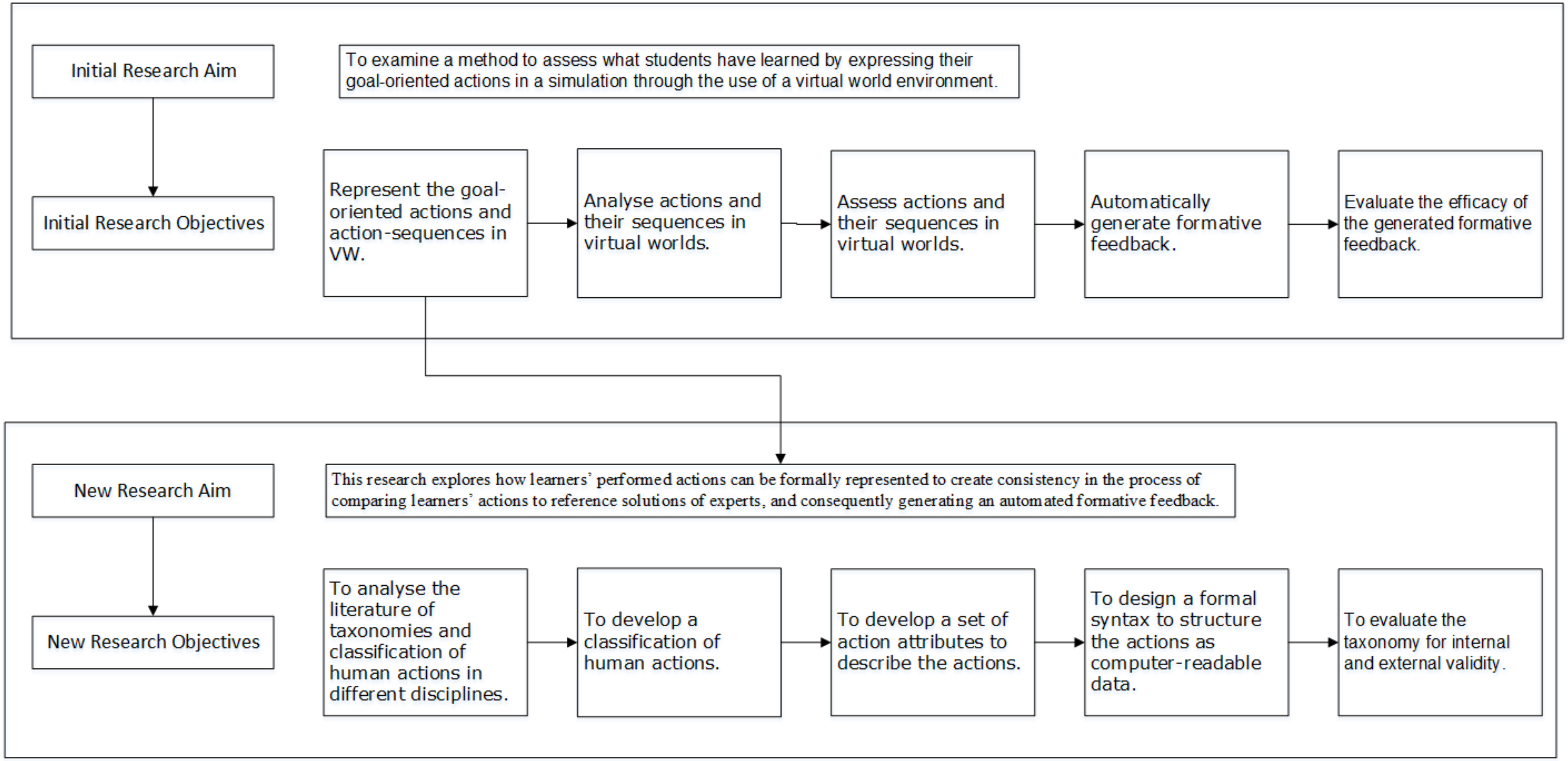

Initial Research Aim
To examine a method to assess what students have learned by expressing their goal-oriented actions in a simulation through the use of a virtual world environment.
Initial Research Objectives
Represent the goal-oriented actions and action-sequences in VW.
Analyse actions and their sequences in virtual worlds.
Assess actions and their sequences in virtual worlds.
Automatically generate formative feedback.
Evaluate the efficacy of the generated formative feedback.
New Research Aim
This research explores how learners' performed actions can be formally represented to create consistency in the process of comparing learners' actions to reference solutions of experts, and consequently generating an automated formative feedback.
New Research Objectives
To analyse the literature of taxonomies and classification of human actions in different disciplines.
To develop a classification of human actions.
To develop a set of action attributes to describe the actions.
To design a formal syntax to structure the actions as computer-readable data.
To evaluate the taxonomy for internal and external validity.

Figure 3.1: Initial and new aim and objectives

3.4. Introduction to Research Design and Methodology

As Chadwick (1984) stated "what is impractical or visionary to one researcher may be pragmatic and utilitarian to another. The important point to remember is that research has little scientific or practical value if it is not properly designed" (p. 27). This research was designed after intensive discussions regarding the problem and its feasible solutions, a preliminary literature review, and the chosen research methodology that was investigated, studied, and tailored to the needs of this study (Section 4.3.2). However, research design is not limited to just carefully choosing the right data and method, but how the selected data and method create new knowledge in a particular area (Chadwick, 1984; Given, 2008; Alturki et al., 2011; Novikov & Novikov, 2013). In this research, various studies on the development of the classification of human actions were investigated, and their evaluation methods were carefully assessed prior to choosing the most appropriate evaluation methods for the created artefact (Section 2.4.5).

3.4.1. Research Paradigms

Mertens (2015) defined the research paradigm as a way of viewing the world, while Creswell (2014) called it 'knowledge claim' and defined it as assumptions made during a project by researchers about what they would learn and how. Collis & Hussey (2014) defined the research paradigm as a framework that "guides how research should be conducted, based on people's philosophies and their assumptions about the world and the nature of knowledge" (p.11). The chosen paradigm helps the researcher to develop the "intent, motivation and expectations" (Mackenzie & Knipe, 2006: p. 2) for the research and, subsequently, the methodology, method and design are selected. After examining various definitions and discussions, this thesis recognises the research paradigm as a framework that is used to develop an understanding of the problem at hand and the way it should be addressed and solved by applying a specific methodology using appropriate methods and tools (Sections 4.5 and 4.6). Although various researchers suggest different paradigms for research, the more common include: Postpositivist (and positivist) (Creswell, 2014; Mackenzie & Knipe, 2006; Collis & Hussey, 2014; Mertens, 2015); Interpretivist/constructivist (Creswell, 2014; Mackenzie & Knipe, 2006; Collis & Hussey, 2014; Mertens, 2015); Transformative (Mackenzie & Knipe, 2006; Mertens, 2015); Pragmatic (Creswell, 2014; Mackenzie

& Knipe, 2006; Mertens, 2015); and Advocacy/Participatory (Creswell, 2014). Design Science Research was used and advanced as an independent research paradigm for long time (livari, 2003, 2007); Hevner et al., 2004; Hevner, 2007; Gregor & Jones, 2007) but started to shift towards a research methodology with the new theoretical advancements (Peffers et al., 2007; Kuechler & Vaishnavi, 2008, 2012; Pries-Heje & Baskerville, 2008). Weber (2010) named three main research paradigms used in Information Systems: (1) interpretive or constructivist paradigm, (2) positivist or postpositivist paradigm, and (3) socio-technical or developmentalist paradigm. Weber (2010) discussed DSR according to its relation to these paradigms and concludes that "DSR is a pluralistic research approach that cannot and should not be separated in an existing research paradigm." (p. 6).

Four basic belief systems (Lincoln et al., 2011; Mertens, 2015: p. 11) that help to define different research paradigms are Axiology (“nature of ethical behaviour”), Ontology (“nature of reality”), Epistemology (“nature of knowledge; relation between knower and would-be known”), and Methodology (“approach to systematic inquiry”). Collis & Hussey (2014) referred to belief systems as assumptions and add Rhetorical assumption (the research language) to the four systems. These four belief systems are addressed by Mertens (2015: p.11) for the four common paradigms:

3.4.1.1. Postpositivist

Axiology: "Respect privacy; informed consent; minimize harm (beneficence); justice/equal opportunity".

Ontology: "One reality; knowable within a specified level of probability."

Epistemology: "Objectivity is important; the researcher manipulates and observes in a dispassionate, objective manner".

Methodology: "Quantitative (primarily); interventionist; decontextualized".

3.4.1.2. Constructivist

Axiology: "Balanced representation of views; raise participants' awareness; community rapport".

Ontology: "Multiple, socially constructed realities".

Epistemology: "Interactive link between researcher and participants; values are made explicit; create findings".

Methodology: "Qualitative (primarily); hermeneutical; dialectical; contextual factors are described".

3.4.1.3. Transformative

Axiology: "Respect for cultural norms; beneficence is defined in terms of the promotion of human rights and increase in social justice; reciprocity".

Ontology: "Rejects cultural relativism; recognizes that various versions of reality are based on social positioning; conscious recognition of consequences of privileging versions of reality".

Epistemology: "Interactive link between researcher and participants; knowledge is socially and historically situated; need to address issues of power and trust".

Methodology: "Qualitative (dialogic), but quantitative and mixed methods can be used; contextual and historical factors are described, especially as they relate to oppression".

3.4.1.4. Pragmatic

Axiology: "Gain knowledge in pursuit of desired ends as influenced by the researcher's values and politics".

Ontology: "Asserts that there is a single reality and that all individuals have their own unique interpretation of reality".

Epistemology: "Relationships in research are determined by what the researcher deems as appropriate to that particular study".

Methodology: "Match methods to specific questions and purposes of research; mixed methods can be used as researcher works back and forth between various approaches".

3.4.1.5. Applied Research Paradigm

Considering the paradigms and belief systems discussed above, this research belongs to Postpositivist (and positivist) paradigm:

Axiology: Participants' privacy is much respected and they are informed of the privacy terms and the signing of a consent form is in order if needed.

Ontology: The research has a specific reality that is within a certain realm of probability.

Epistemology: The researcher manipulates and observes in a dispassionate, objective manner. The participants are independent from the researcher, and they do not affect each other.

Methodology: Design Science Research is used as the research methodology, using mixed methods in its evaluation step.

Rhetorical assumption: “Researcher writes in a formal style and uses the passive voice, accepted quantitative words, and set definitions” (Collis & Hussey, 2014: p. 58).

Table 3.1 summarises the four belief systems for the four common research paradigms. The applied paradigm and belief systems are summarised in Table 3.2.

Table 3.1: The four belief systems addressed by Mertens (2015: p.11) for the four common paradigms

Paradigm	Belief Systems	Explanation

(Post)Positivist	Axiology	Respect privacy;
		Informed consent;
		Minimised harm (beneficence);
		Justice/equal opportunity.

	Ontology	One reality; knowable within a specified level of probability
	Epistemology	Manipulating and observing in a dispassionate, objective manner

	Methodology	Quantitative (primarily);
		Interventionist;
		Decontextualised.

Constructivist	Axiology	Balanced representation of views; Raise participants' awareness; Community rapport.
	Ontology	Multiple, socially constructed realities
	Epistemology	Interactive link between researcher and participants; Values are made explicit; Create findings.

	Methodology	Qualitative (primarily);
		Hermeneutical;
		Dialectical;
		Contextual factors are described.

Transformative	Axiology	Respect for cultural norms; Beneficence is defined in terms of the promotion of human rights and increase in social justice; Reciprocity.
	Ontology	Rejects cultural relativism; Recognises that various versions of reality are based on social positioning; Conscious recognition of consequences of privileging versions of reality.
	Epistemology	Interactive link between researcher and participants; Knowledge is socially and historically situated; Need to address issues of power and trust.
	Methodology	Qualitative (dialogic), but quantitative and mixed methods can be used; Contextual and historical factors are described, especially as they relate to oppression.
Pragmatic	Axiology	Gain knowledge in pursuit of desired ends as influenced by the researcher's values and politics.
	Ontology	Single reality; All individuals have unique interpretation of reality.
	Epistemology	Relationships are determined by what the researcher deems as appropriate to that particular study.
	Methodology	Match methods to specific questions and purposes; Mixed methods.

Table 3.2: The four belief systems for the applied paradigms in this research

Paradigm	Belief Systems	Explanation
Postpositivist	Axiology	Participants' privacy is much respected; Participants are informed of the privacy terms; Participants sign a consent form if needed.
	Ontology	Specific reality that is within a certain realm of probability
	Epistemology	Manipulating and observing in a dispassionate, objective manner; Participants are independent from the researcher; Participants do not affect each other.
	Methodology	Design Science Research, using mixed methods in its evaluation step
	Rhetorical Assumption	Written in formal style, using the passive voice, accepted quantitative words, and set definitions

3.4.2. Research Methodologies

Methodology is inherent to the research process and includes a set of methods (Collis & Hussey, 2014) that specify the techniques for collecting and analysing data. The collectable data might be primary, which originate from an original source, or secondary, which are collected from a currently existing source (Kumar, 2014). The research methods/approaches can be quantitative, qualitative or a combination of both. The choice of methods depends on the research paradigm, research types, and methodology. Research paradigms are discussed in Section 3.4.1. The basic research types are as follows (Kothari, 2004):

- Descriptive vs. Analytical: Descriptive research includes different kinds of surveys and fact-finding studies. The main goal is to describe the current state of affairs. Meanwhile, analytical research uses and analyses facts and/or available information to make a critical evaluation of the state of affairs.
- Applied vs. Fundamental: Applied research is intended to find a solution to an immediate problem; fundamental research is chiefly concerned with generalisations and the formulation of a theory.
- Conceptual vs. Empirical: Conceptual research is concerned with abstract idea(s) or theory. It is usually used to develop new concepts or to reinterpret existing ones. On the other hand, empirical research values only experience or/and observation, often regardless of system and theory.
- Others: There are variations of the abovementioned types.

Kothari (2004) included Quantitative vs. Qualitative as research types although he simultaneously acknowledged them as research approaches. The choice of methodology is based on the research paradigm; and based on the methodology and the research type, the methods/approaches are chosen. As this research is based on the postpositivist paradigm, the methodologies most used in this paradigm, including Experimental studies, Surveys (using primary or secondary data); Cross-sectional studies; Longitudinal studies, and Design Science Research (Collis & Hussey, 2014; Weber, 2010) are investigated. Of these methodologies, Experimental Studies, Surveys, and Design Science Research are the centre of the research presented in this report and are discussed in the following paragraphs.

Experimental studies are conducted to discover relationships between variables by manipulating an independent variable and observing its effect on the dependent variable (Collis & Hussey, 2014). Experimental studies can be in the form of repeated-measure design, independent-sample design, matched-pairs design, or single-subject design (Kothari, 2004; Collis & Hussey, 2014; Creswell, 2014). After the type of experimental design has been decided, the number of groups and the sample sizes should be determined. Kervin (1992) suggests considering three primary factors in experimental studies:

- the number of groups: the comparison can be made between two or more groups, or within one group of participants;
- the nature of the groups: the formation of the groups such as random allocation or matched cases;
- the timing of the experiments: the experiments may be repeated a number of times, or can be limited to one time only.

Surveys are used for the purpose of collecting primary or secondary data from a sample and the data are analysed in order to generalise the findings to a population (Kothari, 2004; Collis & Hussey, 2014; Creswell, 2014). When conducting surveys, the sample size is very important. If the population is small, it is easy to gather the data, but if the population is large, a proper random sample size can be used. Depending on a survey’s purpose, the survey can be descriptive or analytical. If the survey is intended to represent a phenomenon in one or multiple points of time, the descriptive survey is used. Whereas, analytical surveys are used to determine whether there are any relationships among variables. In order to identify the dependent and independent variables, an analytical survey requires a theoretical framework derived from the literature.

The Design Science paradigm, which is a problem-solving paradigm, has been used in the engineering domain for a long time and has been used in the Information Technology (IT) and Information Systems (IS) domains for the past two decades (Venable, 2006). It strives for the creation of new ideas, technological abilities and products whereby the analysis, design, implementation, organisation, and application of information systems can be successfully and competently carried out (Hevner et al., 2004). Design Science strives to "create things that serve human purposes" (March & Smith, 1995: p. 253). The creation of such artefacts depends on current core theories that are exerted, evaluated, and improved through the practice, originality, insight, and problem-solving abilities of the researcher (Hevner et al., 2004). IT and IS artefacts are generally defined as constructs (concepts form the vocabulary of a domain), models (set of propositions or statements expressing relationships among constructs), methods (set of steps, an algorithm or guideline, used to perform a task), instantiations (realization of an artefact in its environment) (March & Smith, 1995), and better theories (Purao, 2002; Rossi and Sein, 2003 cited in Vaishnavi & Keuchler, 2004). These are tangible instructions that empower IT and IS researchers to recognise and address the problems inherent in developing and successfully implementing information systems within organisations (Nunamaker et al., 1991). The result of DSR must be described well, empowering its implementation and application in a suitable field (Hevner et al., 2004).

The process of designing this research in order to answer the question 'what if we could automatically assess learners based on their performance and not a written essay?', led to choosing DSR as a research methodology where the focus is on producing an artefact, herein a taxonomy of human actions, and its evaluation. Moreover, as DSR is the main research methodology, this research benefits from various methods borrowed from experimental and survey methodologies along with other methods to evaluate the artefact. Various DSR frameworks are discussed in the next section.

3.4.3. DSR: Different Frameworks

DSR consists of a sequence of consecutive steps planned in the research design. Numerous DSR studies in the literature propose various frameworks (Nunamaker et al., 1991; March & Smith, 1995; Vaishnavi & Keuchler, 2004; Hevner et al., 2004; Peffers et al., 2006, 2007; Offermann et al., 2009; Alturki et al., 2011). Some of these different sequences of steps are shown in Table 3.3. Hevner et al. (2004) and Peffers (2006, 2007) are predominant in DSR literature, although there are several other DSR frameworks and roadmaps which are mostly based on these two frameworks.

Hevner et al. (2004) use seven guidelines for DSR:

1. Design as an artefact;
2. Problem relevance;
3. Design evaluation;
4. Research contributions;
5. Research rigor;
6. Design as a search process; and
7. Communication of research.

Peffers et al. (2006, 2007) proposed a six-step process for DSR, shown in Figure 3.2:

1. Problem identification and motivation;
2. Objectives of a solution;
3. Design and development;
4. Demonstration;
5. Evaluation; and
6. Communication.

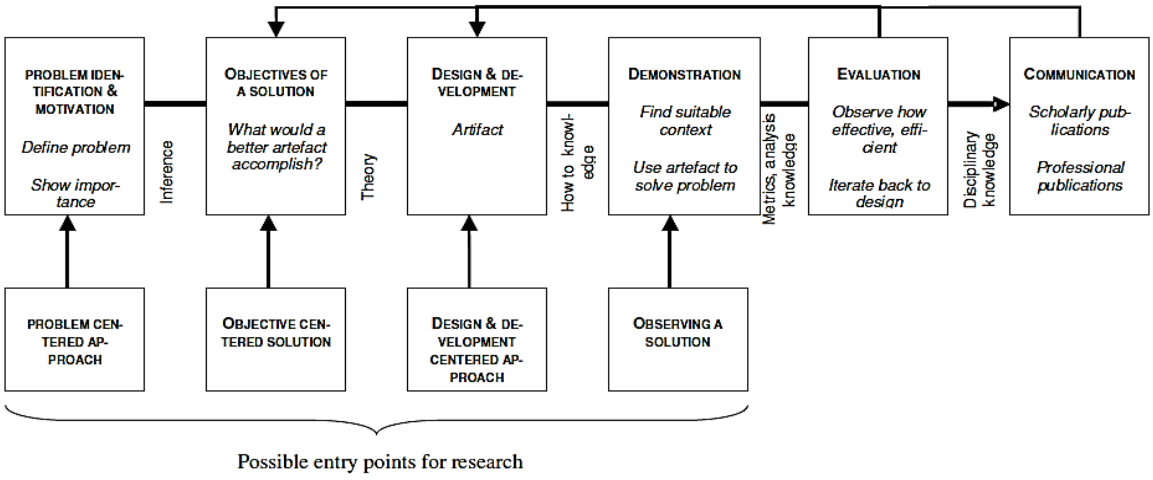

PROBLEM IDENTIFICATION & MOTIVATION
Define problem
Show importance
Inference
OBJECTIVES OF A SOLUTION
What would a better artefact accomplish?
Theory
DESIGN & DEVELOPMENT
Artifact
How to knowledge
DEMONSTRATION
Find suitable context
Use artefact to solve problem
Metrics, analysis knowledge
EVALUATION
Observe how effective, efficient
Iterate back to design
Disciplinary knowledge
COMMUNICATION
Scholarly publications
Professional publications
PROBLEM CENTERED APPROACH
OBJECTIVE CENTERED SOLUTION
DESIGN & DEVELOPMENT CENTERED APPROACH
OBSERVING A SOLUTION
Possible entry points for research

Figure 3.2: Design Science Research process (DSRP) model (Peffers et al., 2006: p.93; with permission for reuse from Professor Ken Peffers)

Offermann et al. (2009) proposed a DSR process which has three main stages: problem identification, solution design, and evaluation. Each stage consists of several detailed steps that can be used based on the needs of the research; these steps are shown in Figure 3.3.

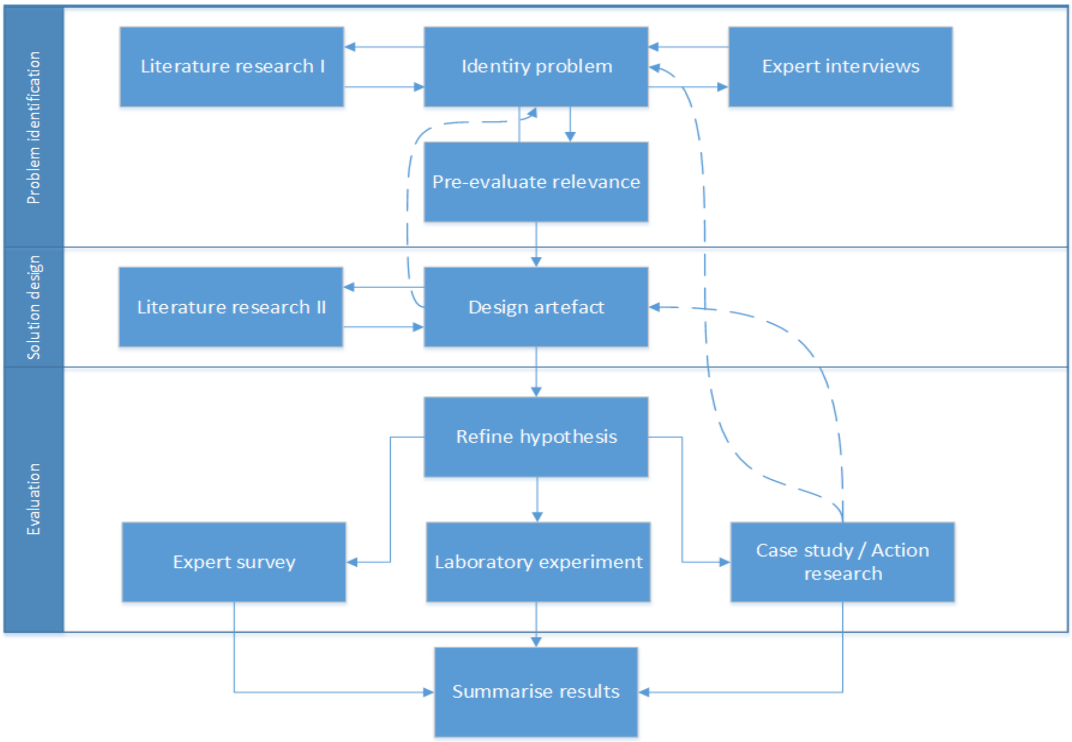


Figure 3.3: DSR process (Offermann et al., 2009)

In contrast to Peffers et al. (2006, 2007), Hevner et al. (2004) offered guidelines rather than a particular DSR process. These researchers recognised models, instantiations, and constructs as an IT/IS artefact, while the previous DSR researches mostly used products as the artefact. Hevner et al. (2004) believed that the main goal of DSR was to solve a problem in a business, and the solution to the problem is an artefact. Peffers et al. (2006, 2007)'s framework begins by identifying a research problem and justifying the value of the proposed solution. Offermann et al. (2009) proposed a literature review and expert interviews for better problem identification and solution analysis. Still, they did not start the design stage before evaluating the relevance of the solution, which is done by creating a research hypothesis and pre-evaluation. In Peffers et al. (2006, 2007), following the problem and solution identification, the aim is to conclude the objectives of a solution of the problem. The framework continues with the design of the artefact as a solution and showing its efficacy in solving the problem. Hevner et al. (2004) believed that artefact design is a constant search for the most satisfactory solution, but not every possible solution, and it involves "the creation, utilization, and assessment of heuristic search strategies" (p. 89). On the other hand, Offermann et al. (2009) suggested another round of literature review which is conducted after artefact design to add the relevant scientific literature. Peffers et al. (2006, 2007) concluded the DSR process by evaluating the artefact and communicating and publishing its findings. Hevner et al. (2004) stated that each design artefact should be evaluated to show its utility, quality, and efficacy.

To evaluate the designed artefact, Hevner et al. (2004) proposed different methods:

1. Observational (case study, field study);
2. Analytical (static analysis, architecture analysis, optimisation, dynamic analysis);
3. Experimental (controlled experiments such as usability);
4. Testing (functional, structural); and
5. Descriptive (informed argument, scenario).

Peffers et al. (2006, 2007) suggested observation (demonstration), and objective quantitative performance measures (satisfaction surveys, client feedback, or simulations) as evaluation methods, while Offermann et al. (2009) used the expert survey, laboratory experimentation, case study and/or action research as an artefact evaluation. Hevner et al. (2004) also emphasised the importance of research rigor and its evaluation; Peffers et al. (2006, 2007) do not evaluate the rigor, but do consider it in the design process. In the final stages of DSR, evaluation and summarisation, Offermann et al. (2009) break down the hypothesis into smaller but more precise parts that, together, support the general research hypothesis/question/aim, and conclude with a summarisation of results.

Table 3.3: DSR steps in literature (steps are in original titles used by each author)

	Nunamaker et al. 1991	March & Smith 1995	Vaishnavi & Keuchler 2004	Hevner et al. 2004	Peffers et al. 2007	Offermann et al. 2009
Problem identification	Construct a conceptual framework		Awareness of problem Suggestion	Design as an artefact Problem relevance Research contributions Design evaluation Research rigor Design as a search process**	Problem identification and motivation	Identify problem Literature research 1 Expert interviews Pre-evaluate relevance
Solution design	Develop a system architecture Analyse and design the system Build the (prototype) system	Build	Development		Objectives of a solution Design and development	Design artefact Literature research 2
Evaluation	Observe and evaluate the system	Evaluate*	Evaluation		Demonstration Evaluation	Refine hypothesis Expert survey Laboratory experiment Case Study / Action Research
Communication			Conclusion	Communication of research	Communication	Summarise results

*March & Smith (1995) consider build and evaluate as DSR and after evaluating the artefact they use ‘theories’ and ‘justify’ as Natural Science Research activities to extract general knowledge by proposing and testing theories.

**These six guidelines apply to all steps in DSR

March and Smith (1995) described artefact evaluation as “the process of determining how well the artefact performs” (p. 254). Evaluation of a designed artefact necessitates suitable metrics and possibly the gathering and analysis of appropriate data (Hevner et al., 2004). A mathematically measurable basis for design enables researchers to use different types of quantitative evaluations of their artefact, including optimization proofs, analytical simulation, and quantitative comparisons with alternative designs (Hevner et al., 2004). However, qualitative evaluation methods are also well established and recognised in DSR evaluation (Vaishnavi & Keuchler, 2004; Peffers et al., 2007; Offermann et al., 2009). As Venable et al. (2012) argued, DSR researchers might use mixed methods depending on evaluation needs.

IT/IS artefacts can be evaluated regarding “functionality, completeness, consistency, accuracy, performance, reliability, usability, fit with the organization, and other relevant quality attributes” (Hevner et al., 2004: p. 85). DSR evaluation can be performed as observational, analytical, experimental, testing, descriptive (Hevner et al., 2004), expert survey, laboratory experiment, case study/action research (Offermann et al., 2009). As Venable et al. (2012) argued, accurate and rigorous, scientific research requires evidence. Moreover, as DSR is claimed to be “science”, then the evaluation must show adequate precision. Hevner et al. (2004), Peffers et al. (2006, 2007), Vaishnavi and Keuchler (2004), Offermann et al. (2009), and other DSR researchers who considered the evaluation phase as one of most important stages of DSR, did not provide very detailed guidelines for evaluation.

Cleven et al. (2009) investigated the prior research on DSR, reference models, and conceptual models (Fettke & Loos, 2003; Pfeiffer & Niehaves, 2005; Frank, 2007; Siau & Rossi, 2011) to create a morphological field of variables and their respective values which are relevant for the evaluation of DSR artefacts. These variables and their respective values are depicted in Table 3.4.

Table 3.4: Variables and values for the evaluation of DSR artefacts (Cleven et al., 2009: p. 3; with permission for reuse from Dr Anne Cleven)

<table>
<tr><th>Variable</th><th colspan="7">Value</th></tr>
<tr><td>Approach</td><td colspan="3">Qualitative</td><td colspan="4">Quantitative</td></tr>
<tr><td>Artefact Focus</td><td>Technical</td><td colspan="3">Organizational</td><td colspan="3">Strategic</td></tr>
<tr><td>Artefact Type</td><td>Construct</td><td colspan="2">Model</td><td>Method</td><td colspan="2">Instantiation</td><td>Theory</td></tr>
<tr><td>Epistemology</td><td colspan="3">Positivism</td><td colspan="4">Interpretivism</td></tr>
<tr><td>Function</td><td>Knowledge Function</td><td colspan="2">Control Function</td><td colspan="2">Development Function</td><td colspan="2">Legitimization Function</td></tr>
<tr><td rowspan="2">Method</td><td>Action research</td><td colspan="2">Case study</td><td colspan="2">Field Experiment</td><td colspan="2">Formal proofs</td></tr>
<tr><td>Controlled experiment</td><td colspan="4">Prototype</td><td colspan="2">Survey</td></tr>
<tr><td>Object</td><td colspan="3">Artefact</td><td colspan="4">Artefact Construction</td></tr>
<tr><td>Ontology</td><td colspan="3">Realism</td><td colspan="4">Nominalism</td></tr>
<tr><td>Perspective</td><td>Economic</td><td colspan="2">Deployment</td><td colspan="2">Engineering</td><td colspan="2">Epistemological</td></tr>
<tr><td>Position</td><td colspan="3">Externally</td><td colspan="4">Internally</td></tr>
<tr><td>Reference Point</td><td colspan="2">Artefact against research gap</td><td colspan="2">Artefact against real world</td><td colspan="3">Research gap against real world</td></tr>
<tr><td>Time</td><td colspan="3">Ex-Ante</td><td colspan="4">Ex-Post</td></tr>
</table>

Vaishnavi & Keuchler (2009: p. 159-172) suggested seven means for the evaluation and validation of the developed solution in DSR: Demonstration ("demonstrate that the solution is realizable and valid in predefined situations"), Experimentation ("to validate or reject a set of hypotheses associated with the claims about the solution"), Simulation ("to evaluate and validate one's solution to the research problem"), Using Metrics ("to aid validation of one's solution to the research problem"), Benchmarking ("to show that one's solution has reasonable performance or is better than some other available solution"), Logical Reasoning ("to argue the validity of the solution"), and Mathematical Proofs ("prove mathematically the claims being made about the solution that one has developed for the research problem"). They strongly suggested that the most robust instrument is mathematical proof, and the least favourable is demonstration. Demonstration is appropriate if the solution is novel and solves a problem for which there is no current solution. In between are logical reasoning or experimentation and simulation. The solidity of logical reasoning depends on the strength and precision of its arguments and assumptions, and it is usually an alternative to experimentation and simulation, which are used when there is a complex problem on hand which is not amenable to mathematical proof. Metrics are useful as a means of quantifying the claims of the solution via mathematical proof, experiments and simulations. If suitable metrics are not available, a weaker alternative would be benchmarking (Vaishnavi & Keuchler, 2009).

In 2012, Peffers et al. investigated the literature pertaining to research that used DSR as the research methodology. Peffers et al. (2012) developed taxonomies of DS artefact types and artefact evaluation methods as results. They classified the artefacts into conceptual and logical artefacts. Conceptual artefacts include “constructs, models, and frameworks, as well as methods, which are conceptual actionable instructions" (p. 401). Logical instructions such as “algorithms and actual hardware or software implementations are classified as instantiations" (p. 401). Artefact evaluation methods are classified into Logical Argument; Expert Evaluation; Technical Experiment; Subject-based Experiment; Action Research; Prototype; Case Study; Illustrative Scenario. Investigating the literature Peffers et al. (2012) mapped the artefacts to the evaluation methods used for each artefact (Table 3.5) that can be used as a guideline when designing new research.

Table 3.5: Mapping artefacts and evaluate methods from the literature (Peffers et al., 2012)

	Logical Argument	Expert Evaluation	Technical Experiment	Subject-Based Experiment	Prototype	Action Research	Case Study	Illustrative Scenario
Algorithm			■					
Construct	▓		▓	░	░			░
Framework								■
Instantiation			■					
Method			■					
Model			■					

Venable et al. (2012) stated five purposes of the evaluation phase in DSR: evaluating artefact’s utility and efficacy (or lack thereof) for achieving its stated purpose; evaluating formalised knowledge about a designed artefact’s utility; comparing to other designed artefacts’ ability to achieve a similar purpose; evaluating for side effects or undesirable consequences of use; and evaluating to identify weaknesses and areas for improvement of an artefact under development. Venable et al. (2012) extended their framework for designing evaluation in DSR (presented in Pries-Heje et al., 2008; Pries-Heje & Baskerville, 2008) by adding a DSR evaluation strategy selection framework, a DSR evaluation method selection framework, and a process or method to use the two extended frameworks. Pries-Heje et al. (2008) and Pries-Heje & Baskerville (2008) try to extend the customary DSR evaluation, ex-post empirical evaluation and add the ex-ante evaluations to broaden the evaluation strategies in DSR. Cleven et al. (2009) also used Pries-Heje et al. (2008)'s contribution and used both ex-post and ex-ante for the timing section of their proposed morphological field for DSR.

3.5. Applied Research Methodology

An investigation of the DSR literature shows that some research studies such as those of Nunamaker et al. (1991), Hevner et al. (2004) and Peffers et al. (2006, 2007) have had the most impact on every other proposed framework for DSR. A review of the extant literature provided a means of focusing the research and a basis for the adoption of a Design Science Research (DSR) (Hevner et al., 2004; Peffers et al., 2006, 2007; Offermann et al., 2009; Venable et al., 2012) methodology. Although the adoption of a specific framework is common practice in most cases, this research noted Venable et al. (2012: p. 427)'s suggestion that the framework be tailored to suit the current research's "resource constraints (e.g. money, equipment, and people's time)". In this research, DSR frameworks developed by Peffers et al. (2006, 2007) and Offermann et al. (2009) were adapted as the main framework and complementary source respectively, and Hevner et al. (2004)'s guidelines were consulted at designing each step because of their positive influence on choice of methods (Peffers et al., 2006). The Peffers et al. (2006, 2007) framework was used because of its flexibility and its generalisability, enabling it to be used in different DS research projects. Although the various steps in the Offermann et al. (2009) framework is used as a complementary source to enrich the adapted framework as it benefits from the various frameworks of its predecessors. The 'Demonstration and Evaluation' step benefits from other research including but not limited to Hevner et al. (2004), Vaishnavi and Keuchler (2004), Peffers et al. (2006, 2007), Offermann et al. (2009), and Venable et al. (2012). Peffer's two steps, demonstration and evaluation, are combined because demonstration is considered as a taxonomy evaluation method. This is also aligned with the Sonnenberg & Brocke (2012) suggestion of mapping these two steps to the evaluation stage, following March & Smith (1995)'s earlier suggestion regarding the build and evaluate stages of DSR. Hence, the two steps are combined to show the validity of the artefact.

The framework consists of five steps (Figure 3.4), beginning with 'Problem Identification and Motivation' and concluding with 'Communication'. In this framework 'Demonstration' and 'Evaluation' are combined as one step because Peffers et al. (2006, 2007) used the demonstration step to show the efficacy of the solution in solving the problem, which is considered as an evaluation method in the literature. These steps are discussed in following sub-sections.

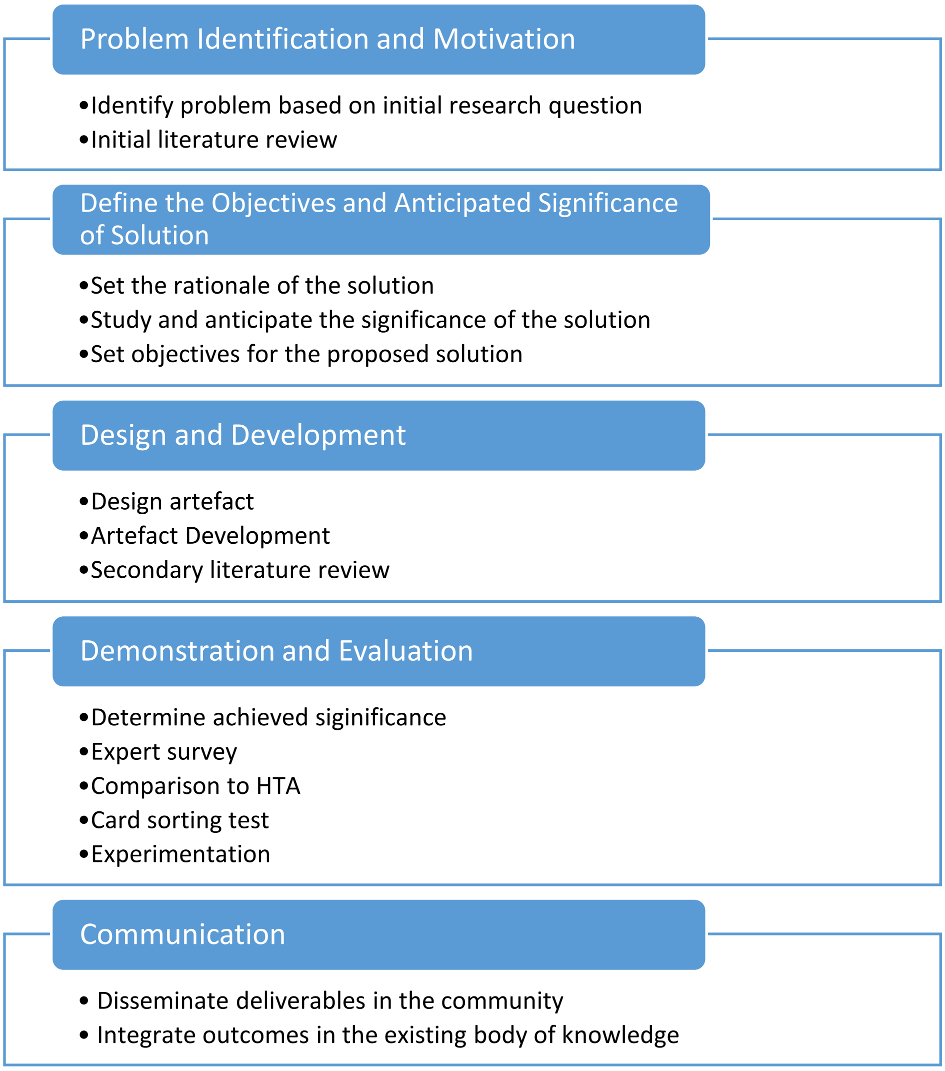

Problem Identification and Motivation
•Identify problem based on initial research question
•Initial literature review
Define the Objectives and Anticipated Significance of Solution
•Set the rationale of the solution
•Study and anticipate the significance of the solution
•Set objectives for the proposed solution
Design and Development
•Design artefact
•Artefact Development
•Secondary literature review
Demonstration and Evaluation
•Determine achieved siginificance
•Expert survey
•Comparison to HTA
•Card sorting test
•Experimentation
Communication
• Disseminate deliverables in the community
• Integrate outcomes in the existing body of knowledge

Figure 3.4: Research DSR framework developed for this research

3.5.1. Step 1: Problem Identification and Motivation

During the process of developing the research proposal, including the initial literature review and discussions, the initial research question was formed. Investigating the answer to the research question led to the identification of the problem. Following the literature investigation, practical answers to the research question were explored, thereby helping to refine the problem. Various feasible solutions were discovered in the course of solving this problem.

During the discussions on automated assessment, a question arose concerning whether learners' performed actions can substitute for written essays and how these actions can be processed and assessed in order to produce an automated formative feedback. Consequently, the literature of learning and assessment, virtual worlds and (2D/3D) VTEs were investigated. As a result, the research problem was defined, and feasible solutions were investigated. In the preliminary investigations, it became evident that, to date, there is no an assessment method which processes the performed actions in (2D/3D) VTE and creates a formative feedback.

After the first round of Design & Development (ALAM: Chapter 4) and the crucial need for a standard representation of the performed actions, the 'Problem Identification and Motivation' step was repeated to redefine the problem. Consequently, following an extensive literature review and discussions with experts, it was discovered that there was no appropriate, exhaustive standard taxonomy of human actions.

3.5.2. Step 2: Define the Objectives and Anticipated Significance of Solution

Feasible solutions were defined (and re-defined) based on literature studies (Chapter 2) and preliminary expert consolidations. In an iterative process, the rationale for the feasible solutions was evaluated, refined and compared to similar problems/approaches in the literature. The most satisfactory solution was selected for the next step. In order to achieve their objectives, experts thoroughly evaluate and assess the significance of the most satisfactory solution. Hence, the research aim was to find a solution that was significant and acceptable. To achieve this aim, the objectives were formulated to solve the problem. The outcome of this stage was the starting point for the design and development process; that is, the "blueprint" for creating the artefact.

3.5.3. Step 3: Design and Development

The outcome of this step was the design and development of an artefact according to the solution defined in the previous step. In an iterative process, in each stage of design, formative artefact evaluation was used to redefine the design until satisfactory results emerged. The artefact was developed based on a defined design to fulfil the research aim. The final artefact was verified and validated against the literature to confirm its state-of-the-art. This step was performed twice as at the end of the first round; new problems emerged that led to re-defining the research problem and repeating steps one to three. Figure 3.5 illustrates the DSR process used in this research.

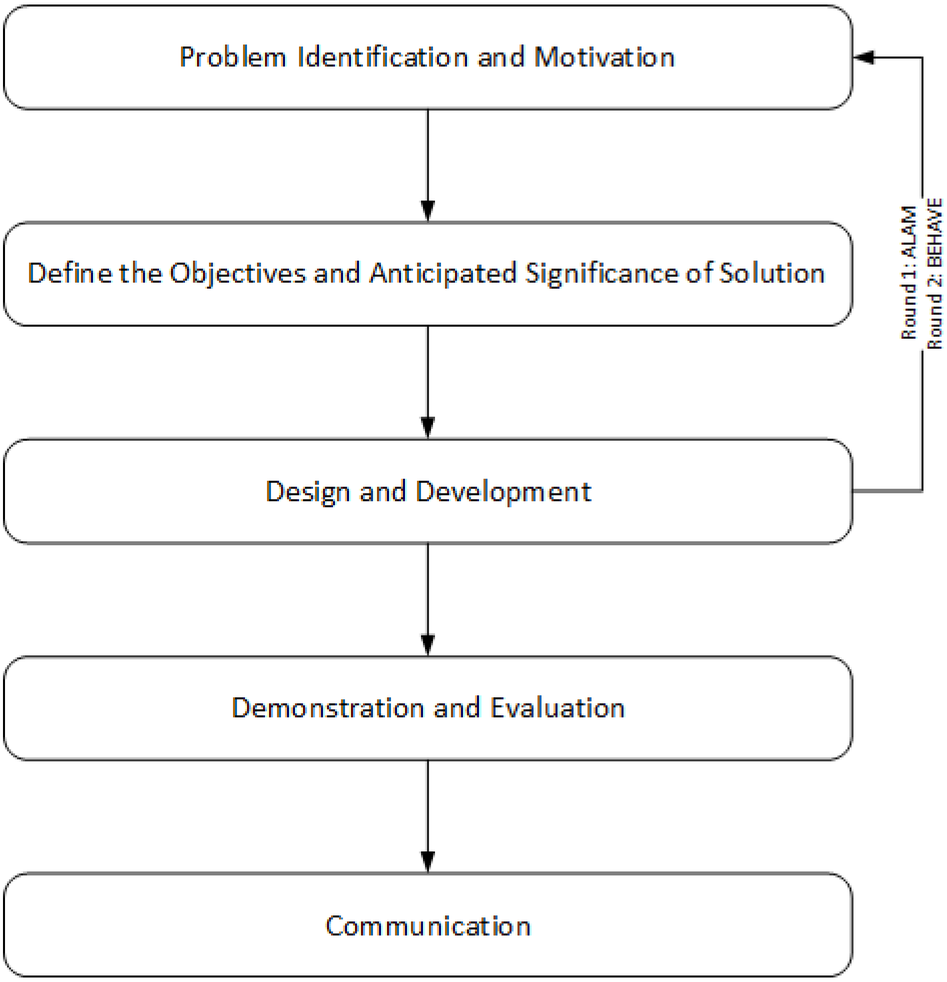

Problem Identification and Motivation
Define the Objectives and Anticipated Significance of Solution
Design and Development
Demonstration and Evaluation
Communication
Round 1: ALAM
Round 2: BEHAVE

Figure 3.5: DRS process used in this research

The first round of the development process led to the ALAM framework explained in Chapter 4 and thence to re-defining the research problem. The second round of the design process started with an investigation of various human actions theories and taxonomies in different disciplines. Then, various real-life and 3D virtual training scenarios and environments were observed. Different levels and classes of human actions were developed. An attributes set was developed to describe performed human actions, and a specific syntax was designed to regulate the use of classification in a codification process of performed actions. Finally, an intensive secondary literature review was conducted in order to compare the developed taxonomy with other available taxonomies to show the state-of-the-art of the former.

3.5.4. Step 4: Demonstration and Evaluation

In this step, the validity of the developed artefact as a solution to the research problem was demonstrated. Several means of evaluation including quantitative or/and qualitative methods were used to represent the internal and external validity of the artefact. Surveys were used for validity evaluation and formative evaluation during the design process. The internal and external validity of the artefact was also evaluated using a card sorting test and experimentation. The evaluation results and analysis are presented in Chapter 6, and discussed in detail in Chapter 7.

3.5.5. Step 5: Communication

As Peffers et al. (2006, 2007) stated, in this step the researcher should "communicate the problem and its importance, the artefact, its utility and novelty, the rigor of its design, and its effectiveness to researchers and other relevant audiences, such as practicing professionals, when appropriate" (p. 92). The dissemination of deliverables in the community occurs via presentations, seminars, and scholarly publications including conference papers, journal articles, books, and reports (Offermann et al., 2009). The various communication methods integrate the outcomes with the existing body of knowledge.

In this research, the researcher used published scholarly papers to communicate with the academic and industrial sectors (p. VI); and finally, the current thesis reports on the research results. The research outputs were presented to practising professionals at various seminars and conferences including, but not limited to, Teaching and Learning Forum (January 2014), National Centre for Research on Evaluation, Standards, and Student Testing (CRESST) - UCLA (March 2014), Three Minutes Thesis Competition (September 2014).

3.6. Evaluation and Data Analysis Methods

As Hevner et al. (2004) state, the gathering of useful data in the evaluation step leads to mathematical evaluation and proof that is, as Vaishnavi & Keuchler (2009) state, the strongest proof in artefact evaluation. Before using a measurement instrument, it is crucial that the researcher be relatively confident that the instrument is both valid and reliable. Validity and reliability are technical characteristics of the measurement instrument.

Validity is “the most critical criterion and indicates the degree to which an instrument measures what it is supposed to measure. In other words, validity is the extent to which differences found with a measuring instrument reflect true differences among those being tested” (Kothari, 2004: p. 73). That is to say, validity tells us whether we are actually measuring a particular concept. Validity means that the measurement instrument shows the true reality; that is, the researcher should ensure that the measurement instruments, including questions and other content, will not gather more information than needed nor omit any necessary information, by measuring the variables. The goal of testing validity is to determine whether the measurement instrument can measure the required characteristics. Without a valid instrument, the accuracy of the information cannot be trusted. One type of validity which is also used in this research is content validity. Content validity is “the extent to which a measuring instrument provides adequate coverage of the topic under study” (Kothari, 2004: p. 74). Content validity can be determined by a group of people who determine whether the measurement instrument adequately covers all the standards. However, there is no numerical way to show the content validity (Kothari, 2004). In this research, the measurement instrument was reviewed and judged by the supervision panel.

The reliability test is an important test to determine the soundness of the measurement instrument. “A measuring instrument is reliable if it provides consistent results” (Kothari, 2004: p. 74). The reliability coefficient varies within a range of 0 (Non-Reliable) to +1 (Reliable). The reliability coefficient indicates the extent to which the instrument measures the stable or interim characteristics. To measure the reliability coefficient of the measurement instrument, various methods are used including Test–Retest, Equivalence, Split–half, Kuder–Richardson, and Cronbach’s Alpha. In this research, the Cronbach’s Alpha is used to determine the reliability. Variance analysis is used to measure the reliability coefficient. The following formula is used to calculate the Cronbach’s Alpha in which N is number of questions, σ_x^2, variance of total questions of the questionnaire, and $\sum_{i=1}^{N} \sigma_{Y_i}^2$, total variance of questions 1 to N:

$$Alpha = \left(\frac{N}{N-1}\right)\frac{(\sigma_X^2 - \sum_{i=1}^{N}\sigma_{Y_i}^2)}{\sigma_X^2}$$

George and Mallery (2013) provided a rule of thumb for Cronbach’s Alpha reliability coefficient stating that “> 0.9 – Excellent, > 0.8 – Good, > 0.7 – Acceptable, > 0.6 – Questionable, > 0.5 – Poor, and < 0.5 – Unacceptable” (p. 231).

Various factors are taken into account when choosing an appropriate statistical analysis method. These include the number of groups, the number of participants, and type of data required. An appropriate method can be chosen by using a decision tree as in Figure 3.6.

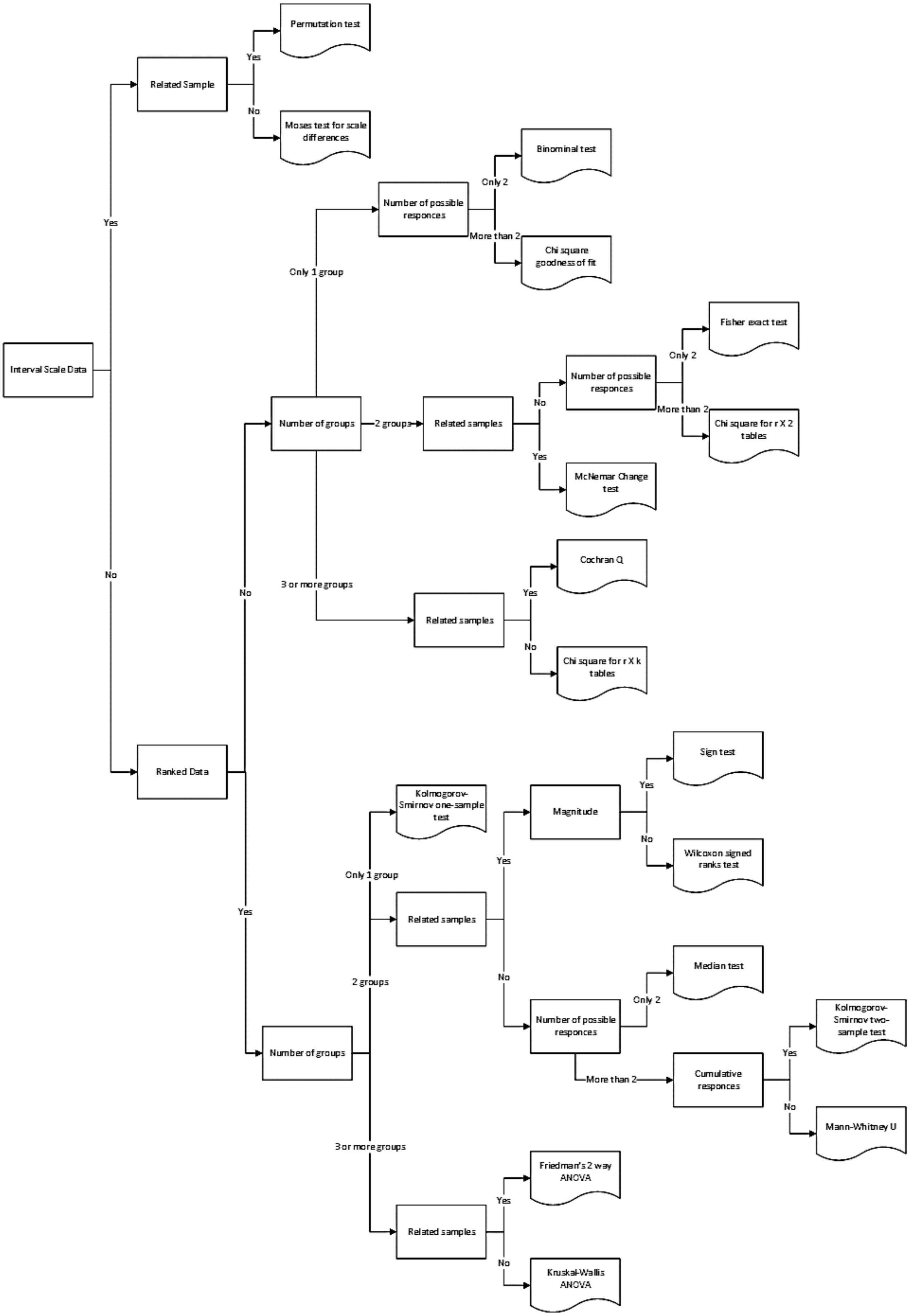


Figure 3.6: Choosing appropriate statistic tests (Kothari, 2004)

3.6.1. Survey for Expert Opinion: Virtual Worlds and Industry Experts

Following the researcher's study of human actions taxonomies and classification, and after the levels and classes of BEHAVE had been developed, a survey was conducted among experts in VW and industry to seek their opinions about the importance of each level and class in the classification section of BEHAVE. The respondents were experts from the manufacturing and production sector, and members of the virtual world working group (VWWG) (Section 3.7). The survey consisted of three main sections including the importance levels and classes of BEHAVE, the importance of different key characteristics of ALAM, and an open question to introduce any known taxonomy of human actions (Appendix 1). The results were analysed using appropriate statistical methods (Sections 6.2 and 7.2). The groups of respondents are shown as G, and the survey questions are shown as Q to form the hypothesis, accompanied by the variables i and j to represent the number of groups or questions.

Chi-square is a non-parametric test which can be used as a test of goodness of fit or/and as a test of homogeneity. Chi-square is used as a test of goodness of fit in order to determine whether the hypothesis fits the observed data (Kothari, 2004). In this research, Chi-square is used to show the fit of the hypothesis which assumes that respondents have a preference in choice towards the importance of the BEHAVE levels and classes, to the observed data.

Friedman's test is used to prioritise the BEHAVE levels and classes based on respondents' opinion. The Binomial Test is used as an alternative to the Chi-square test because the respondents had only two options to choose from, instead of the five options in other questions. Finally, because the contingency table is 2*2, and the observed frequency of two cells is less than five and cannot be merged, the p-value obtained from the Exact Fisher's test is used instead of Chi-square statistics as a test of goodness of fit. Chi-square is also used to compare experts' opinions in the survey, as a test of homogeneity.

3.6.2. Card Sorting Test

Card sorting is a common practice to "elicit end user input into the organization of an information structure" (Hannah, 2008: p. 4). Although card sorting is known to be a method for the design and evaluation of website architecture, accompanied by cluster analysis, it can be used to investigate clusters in taxonomies. It can be performed both manually and by computer (Spencer, 2009). Card sorting test can help the researcher to study the way that people use the taxonomy, and recognise the problems in the clustering of sorted items. The card sorting test helps to improve and at the same time show the validity of the classes under which items were sorted. In this research, the computerised web-based card sorting test was used with participants via the Find Participants website (Section 3.7). K-mean cluster analysis, R-square, and Pearson correlation test were used for data analysis of the test which confirmed the exclusiveness of the classes (Sections 6.3 and 7.3).

In this research, the card sorting test is used for taxonomy validation. The clusters are chosen based on BEHAVE Functional classes of actions, and 47 random actions are chosen to avoid the risk of low test validity. The 47 actions are chosen by asking several people to suggest a list of actions they might perform in life. The lists were compared to avoid duplication. The result of the process led to 47 actions that were used as cards in the card sorting test. Various statistical analysis tests are used to investigate different aspects of taxonomy validity, such as exhaustive clusters, correlations, and dispersion of data inside the clusters.

Cluster analysis is principally used to discover clusters in data. Methods used for clustering should not be confused with ‘discrimination and assignment methods’, “where the groups are known a priori, and the aim of the analysis is to construct rules for classifying new individuals into one or other of the known groups” (Everitt et al., 2011: p. 7).

The K-mean cluster analysis method is the most practical data clustering method. In this method, the number of clusters is fixed and predetermined. It is designed for clustering data numerically (quantitative) and the cluster has a centre called ‘mean’ which is the average of all the data points in the cluster. In this method, the n data points are partitioned into k clusters. The K-mean clustering process is as follows:

1. After determining the number of clusters, the centre of each cluster is initialised.
2. Each data point is attributed to the cluster with closest centre distance to the data point.

3. Each cluster’s position is set to the mean of all data points in the cluster.
4. This process repeats until convergence (minimizing the within-cluster sum of squares).

The Pearson correlation test (Pearson, 1895) is used to investigate the correlation between the taxonomy classes which contributes to the taxonomy validity. Also, the coefficient of determination or R-squared (Nagelkerke, 1991), which specifies how well data fit a statistical model, is used to determine the internal validity of partitioning. The Fowlkes and Mallows index (Fowlkes & Mallows, 1983), which is an external evaluation method, is used for measuring the similarity between clusters.

3.6.3. Performance Coding Experiment and Participant Feedback

Experimentation is a well-known evaluation method used in DSR. In this research, an experiment was developed to evaluate the applicability of the taxonomy and study the degree of similarity between the coded performance and the real-life performance. The participants (Section 3.6) were given a scenario and an online coding tool, and were asked to use BEHAVE (Classes, attributes) and code the scenario. The online coding tool created the syntax and the coded performance. Following the experiment, participants were asked to provide feedback on the degree of similarity between the coded actions and real-life actions. The results can be found in Sections 6.4 and 7.4.

For a computerised assessment of human actions, it is crucial to provide very precise information on performed actions to present the actions in code that is as close as possible to the real-life performance. Moreover, increasing such possibility leads to the increase of quality of the assessment. The rationale behind this validity test is based on the need for a high degree of similarity between performed actions and the taxonomic codification of those actions.

Two groups of participants were asked to use an online coding tool to code a performed scenario using BEHAVE. The first group were given a written scenario and the second group viewed a recorded video of the same scenario. At the end, both groups were asked to provide an answer to two open-ended questions:

What do you think about the provided classes of actions and their attributes?

How close could you code the performance, using the provided taxonomy of human actions and given action attributes, compared to a real-life situation?

Although the main goal was to measure the degree of similarity based on experts' opinions, investigating the experts' performance in the experiment also shed some light on different human factors. The results are presented in the following sections.

The Student's t-test is a significance test based on t distribution to apply significance contribution applicable to small sample groups. Two conditions must exist in order to be able to use a t-test: first, the sample population under 30 and second, unknown population distribution. There are certain assumptions when using t-test including normal or approximately normal sampling population; sample being random; independent observations; no measurement error and equal population variances when testing the equality of the two population means (Kothari, 2004).

In this experiment, in order to study the answers given by both groups of participants, video and written narration, one-sample Student's t-test is used for inference on population's mean. The decision-making chart (Figure 3.7) for choosing the best mean comparison test indicates why the Student's t-test was chosen. The Kolmogorov-Smirnov normality test is used before the t-test to check whether the variables' distribution is normal, and therefore, whether or not we can use the t-test.

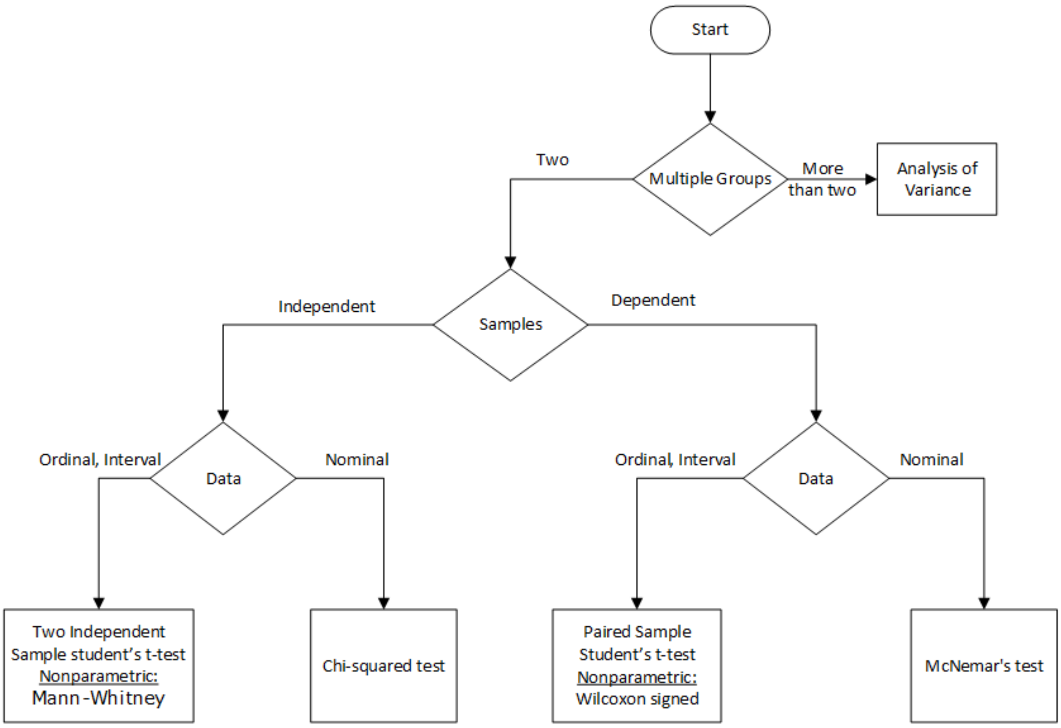


Figure 3.7: Decision-making chart for choosing the best mean comparison test

Table 3.6 summarises the evaluation methods and analysis tools.

Table 3.6: Evaluation methods and data analysis tools

Evaluation Method	Analysis Tools
Expert Opinion Survey	• Cronbach’s Alpha reliability coefficient • Friedman's test for prioritising • Chi-square goodness of fit • Binomial test • Chi-Square Test of Homogeneity
Card Sorting Test	• Scatterplot matrix • K-mean Cluster analysis • Pearson correlation test • Fowlkes and Mallows index • R-squared
Performance Coding Experiment	• One-sample t-test • Two independent sample t-test • Kolmogorov-Smirnov normality test

3.7. Participants

The participants for the different phases of the research were chosen from various populations based on the purpose of the study and the expertise required of participants. The participants were approached in various ways including email distribution and online participant targeting services. For three different studies conducted in this research, different participants were invited based on the goal of each study. The participants invited to undertake the survey study were experts who had experience in education (especially training) in both simulated and real working environments. For the card sorting study, the participants were targeted by specific demographic characteristics as no particular expertise was needed. For the coding experiment, participants with, but not limited to, engineering (especially mechanical engineering) or/and education backgrounds were encouraged to participate. More detailed information on the participants for each study is provided in the following paragraphs.

For the survey, two groups of participants were approached via emails: manufacturing and production experts, and the virtual world working group (VWWG) members (http://www.vwwg.info). Although online surveys have showed lower response rates (20%-40%), in different research over the years because of the vast reduction in time and cost of the surveys, this method is very popular among researchers (Smee & Brennan, 2000; Shannon et al., 2002; Shannon & Bradshaw, 2002; Nulty, 2008). In this research, judgmental sampling as a type of non-probability sampling is used, as this method provides more reliable results compared to random sampling (Levy & Lemeshow, 2013). The population of industry experts comprised

50 engineer experts, and the VWWG group had 80 active members with available contact detail on Wikispaces group, at the time of this research. As for the VWWG, an invitation to participate was sent to all 80 members. Twelve of these addresses were not deliverable at the time, and only 37 emails were reported as opened. Of the latter, 23 experts agreed to participate and 18 of them submitted their answers to the survey on the Qualtrics surveying website. The return rate of 48% is an acceptable rate. The invitation sent to a manufacturing company was forwarded to engineers and technicians by the supervising manager. After the distribution of invitations, two more follow-up emails were sent out.

For the card sorting study, emails were sent to the participants using Find Participants (https://www.findparticipants.com/) services. Emails were sent by the system to 948 potential participants, followed by three reminders. The participant pool was narrowed down by demographic characteristics including age of 18 and above, education level of Diploma and above, and English language. Two hundred and fifty-five respondents opened the test link, and 207 of these completed the test and submitted responses.

The call for participation in the performance coding experiment was advertised at Curtin University, among students and staff of different departments, especially engineering and education. As this method of recruitment aims to attract participants openly, the population is unknown. Although there are no strict guidelines for non-probabilistic sample size (Guest et al., 2006; Francis et al., 2010), researchers such as Manson (2010) investigated the literature and research to identify the best examples of non-probabilistic sample size selection. Various sample sizes are proposed in the literature (Kuzel, 1992; Morse, 1994; Creswell, 2014; Bernard, 2000). The initial sampling size of seven was selected based on Guest et al. (2006), and following the Francis et al. (2010) method, a set of three participants was added as a stopping criterion. This group was given a written narration to code and then a semi-structured interview was conducted. Since the data saturation began to emerge before the sixth interview and by the next set of three interviews added to the initial group, it was decided to investigate whether a greater degree of freedom would change the outcome. Consequently, another group of ten participants, a set of three in addition to the initial, were presented with a video of the same scenario instead of the written narration.

Despite the addition of new participants, very few new points emerged during the interviews, thereby indicating a data saturation in the second group as well.

3.8. Research Ethics

This study was carried out within the guidelines of the NHMRC National Statement on Ethical Conduct in Human Research:

(http: //www.nhmrc.gov.au/publications/synopses/e72syn.htm)

Collected data has been kept confidential, and only the researcher and named supervisors have access. Participants in this study remain anonymous and unidentifiable in any published material. There is no risk for participants, and they were informed that their participation in the study was completely voluntary. They were free to withdraw at any time without explanation.

3.9. Summary

In this chapter, the research aim and objectives were discussed. The research design and methodology and the reason for choosing DSR as the research methodology were presented. Furthermore, the evaluation methods, data analysis methods, participants, and research ethics were introduced. In the next chapter, Chapter 4, the Action-based Learning Assessment Methodology (ALAM) framework is explained.

Chapter 4: Action-based Learning Assessment Methodology

4.1. Introduction

This chapter explains the Action-based Learning Assessment Methodology (ALAM) framework, its development process, characteristics, and the conceptual model for an assessment system based on ALAM. The primary objective of ALAM is to provide an automatically generated detailed post-performance formative feedback to learners based on their performed actions. As illustrated in Figure 3.5, ALAM is the artefact developed during the first round of DSR methodology in the initial stage of this research.

The learner's actions are performed for the purpose of solving a problem, completing a scenario, or fulfilling other possible expectations that show that the learner can apply the memorised knowledge in practice, during the assessment. The given problems or scenarios may be chosen from real-world settings; for this matter, Section 4.2.1 discusses how ALAM supports authentic assessment.

Furthermore, the ALAM framework is introduced (Section 4.3) that consists of three main components, each in charge of a particular part of the assessment process. The first component is the Performance Codifier Engine (PCE). After the codification of the learner's performance and experts' reference solutions, the Comparison Engine (CE) will compare the coded performances. Then, CE maps them together based on given rules, set by the experts, so the Feedback Compiler Engine (FCE) can use the results to create a detailed formative feedback and send it to the learner.

The development process of ALAM can be summarised in the following steps:

1. Reviewing the literature on various theories and methods of Action-based Learning, formative feedback, and automated assessment (Sections 2.2, 2.3, and 2.4);
2. Determining the main characteristics (Section 4.3);
3. Developing the ALAM components (Section 4.3);

4. Developing the conceptual model of an assessment system based on ALAM (Section 4.4).

4.2. Action-based Learning Assessment Methodology

ALAM is a new assessment methodology that automatically generates a post-performance formative feedback for learners based on their actions and action-sequences performed in a (2D/3D) VTE as the assessment environment. In brief, ALAM receives the information on the performed actions from the (2D/3D) VTE directly or via a third-party technology, and then maps the actions, and the descriptions of actions and sequences to a formalised coding syntax (Section 4.3.1). Finally, ALAM compares these coded actions to multiple reference solutions created by experts (Sections 4.3.2) and generates an automated formative feedback (Section 4.3.3).

ALAM adapts the terms ‘Outcome goals’, ‘Performance goals’, and ‘Process goals’ from the Sport and Exercise Psychology and redefines them according to its needs. Weinberg & Gould (2014) suggested the following definitions: Outcome goals "typically focus on a competitive result of an event, such as winning a race, carving a medal, or scoring more points than an opponent" (p. 352). Performance goals "focus on achieving standards or performance objectives independently of other competitors" (p. 352). Process goals "focus on the actions an individual must engage in during performance to execute or perform well" (p. 352).

On the other hand, ALAM redefines the terms as follows:

Outcome goal is the final and main goal set for the learner to achieve during the assessment. The Outcome goal may be a problem to solve, a scenario to be completed, or other possible expectations to be fulfilled.

Performance goals are strategically chosen by the learner based on prior knowledge, and must be achieved in order for the Outcome goal to be realised.

Process goals are realised by the actions the learner must perform during the performance in order to achieve a successful Performance goal.

Assessment of action choices is used in educational games and VTEs for summative and formative assessment of memorised knowledge and, in some cases, application of the learned knowledge (Shute et al., 2009; Chodos et al., 2014; Shute &

Ventura, 2013). However, ALAM analyses and assesses how learners do things, and not just what they do, although the action choices are still part of the assessment process.

The main difference between ALAM, as an assessment method, and other similar assessment methods involving learners' performance, is that ALAM does not restrict the learner with predefined action choices as do most assessment methods in educational games. Learners undertake the full performance, and they see the consequences of their actions within the limitations of the designed system. Of course, these limitations should be addressed and minimised by the VTE assessment scenario developers by predicting different consequences of each probable action, to possible extends. However, since the (2D/3D) VTE has been developed for learning and assessment purposes, one cannot expect the developers to consider limitless probable choices of actions, as learners might choose (for any reason) to perform a sequence of actions completely irrelevant to the assessment scenario that could not be foreseen by the developers. Even in the case of complete irrelevancy, ALAM recognises the irrelevant actions and reflects on them in the generated feedback.

Based on performed actions and the sequences of those actions, a formative feedback is generated by ALAM that evaluates the learner's performance by identifying the possible mistakes (based on comparison to the reference solutions) and best given solution. Consequently, learners can learn from the provided feedback and correct their mistakes or improve their performance. Learners can master the needed skills for real-life performances by applying the formative feedback generated by ALAM. This level of learning is at the application level of Bloom's taxonomy (Bloom et al., 1956).

Moreover, ALAM incorporates the 'eight critical elements of authentic assessment' as proposed by Ashford-Rowe et al. (2014). The support for authentic assessment is aligned with several aspects considered in the development of ALAM such as (but not limited to) being based on actions and performance, formative feedback, and learning from the assessment. The following sub-section studies these aspects.

4.2.1. ALAM Supports Authentic Assessment

ALAM allows learners to learn from their mistakes by performing in a simulated environment in order to master the required skill in real-life situations such as the workplace. Whitlock & Nanavati (2013) presented authentic assessment as the most suitable assessment method in situations where the learner is expected to apply, analyse, evaluate, or create. Students can benefit from authentic assessment in simulated or real-world contexts as it enables them to develop the confidence "to successfully accomplish those tasks on their own in subsequent, similar situations" (Whitlock & Nanavati, 2013: p. 36).

ALAM supports the authentic learning activity criteria set by Ashford-Rowe et al. (2014) in their framework for designing authentic assessment. These eight criteria are discussed in the following paragraphs:

1. "An authentic assessment should be challenging":

The assessment scenarios and given problems are real-world scenarios based on the course the learners are assessed on. The learners are challenged the same as they would be in real-life scenarios during the assessment. They are asked to analyse the given task or problem and use the skills they have acquired during the course to choose the most appropriate response.

Each scenario in ALAM consists of an overall goal known as the outcome goal, a number of milestones known as the performance goals, and a number of actions constituted each milestone, known as process goals. Learners need to use their previous knowledge to create new solutions and apply it to the problem at hand by choosing the milestones and actions that are most appropriate for achieving the main goal or solving the problem.

2. "The outcome of an authentic assessment should be in the form of a performance or product (outcome)":

To fulfil this criterion, the learner should demonstrate skill by the application of knowledge during the assessment. Although the main outcome of ALAM is formative feedback, which may be considered as a product, learners' skill is reflected by their performance that consists of a sequence of actions chosen by the learner to achieve the given goal or solve the given problem. Therefore, the performed actions are considered as outputs of the assessment that fulfil the criterion for the authenticity of this assessment method. However, depending on the chosen scenario, an additional output might exist in the form of a product that is assessed in the process based on the pre-set characteristics defined by the experts.

3. “Authentic assessment design should ensure transfer of knowledge”:

Authentic assessment should ensure that the learner can apply the learned knowledge in other domains as well. However, although this element is not directly satisfied by using ALAM as an assessment method, based on the structure of the assessment scenario or series of scenarios, this element can be satisfied. As ALAM is independent of the field of the assessed knowledge, the transfer of knowledge should be facilitated by the curriculum design. However, if ALAM is used with a bottom-up approach, learners can master isolated tasks (a sequence of actions to achieve a certain goal) that would be used in various situations in addition to other tasks.

4. “Metacognition as a component of authentic assessment”:

Authentic assessment should enable learners to learn from their assessment. Detailed constructive feedback can help learners to reflect on their knowledge and skills. The main output of ALAM is a detailed formative feedback that enables learners to learn from their assessed performance. This learning can occur at different levels based on the type of feedback that is provided and the information it delivers. In most cases, learners do not only receive information about their mistakes, but also learn how to correct them.

Moreover, a feedback can provide suggestions, corrections, or/and improvements that learners can use to improve their performance and consequently their skills. This information might be in the form of best practices, methods different from those used by the learner, and suggestions regarding alternative or additional actions.

5. “The importance of a requirement to ensure accuracy in assessment performance”:

The aim of authentic assessment is not just to show whether a goal has been achieved; it is also intended to reveal the process leading to this achievement. Facilitated by (2D/3D) VTE, ALAM requires learners to demonstrate the application of their knowledge by performing in a simulated real-life scenario. Furthermore, ALAM assesses the similarity to expert performances by analysing each performed action, action-description, and action-sequence. Analysing each action performed to achieve the assessment goal enables ALAM to conclude whether the process of goal achievement was accurate. The analysis of the performed actions and its comparison to reference solutions reflect the extent to which the learner's skills are comparable to real-life applications.

6. "The role of the assessment environment and the tools used to deliver the assessment task":

Authentic assessment should provide an assessment environment that closely approximates the real-world settings for the learner. However, sometimes the re-creation of the real world is difficult to achieve in a simulated training and assessment environment. Culturally familiar settings also need to be set in place, by means of language and familiar images.

Another important ALAM criterion is the assessment environment. For the assessment to be as authentic as possible, ALAM suggests that (2D/3D) VTE be used as an assessment environment. Nevertheless, various other factors such as the degree of immersion and authenticity of the environment should be considered. Depending on the scenario, a 2D simulation (e.g. machine control panel) might be regarded as highly authentic, although another scenario might require an expensive simulator cabin (e.g. flight simulators).

7. "The importance of formally designing in an opportunity to discuss and provide feedback":

As discussed under Criterion 4, learners need to reflect on their performance. This reflection is facilitated by a formative feedback provided to the learner at the end of the assessment. ALAM provides a detailed formative feedback on learner's performance that helps learners to learn from their mistakes and improve the required skills. Although ALAM provides the feedback, there is no subsequent discussion unless this occurs in a classroom setting, whether it be a physical classroom or online classroom.

8. "The value of collaboration":

Authentic assessment should allow the learner to communicate with others during the assessment process. With ALAM, collaboration is possible as each recorded action in the environment is tagged with the learners' ID. Hence, during the action mapping process, team members' actions can be mapped and assessed individually or in respect to others.

However, in the case of collaboration and teamwork, usually, the teamwork should be analysed and assessed as a group effort, r and not be based on individual performances. This option can be embedded into the ALAM assessment mechanism by mapping the collaboration and creating special rules and relations for collaborations (Section 5.2.2).

4.2.2. ALAM Development Process

ALAM was developed using a four-step process:

1. Reviewing the literature on various theories and methods of Action-based Learning, formative feedback, and automated assessment (Sections 2.2, 2.3, and 2.4);

During the 'Problem Identification and Motivation' step of applied DSR methodology, an initial literature review was conducted on learning and assessment, various Action-based Learning theories and methods (e.g. experiential learning, problem-based learning, authentic learning, and scenario-based learning), automated assessment, and learning and assessment in simulated environments.

2. Determining the main characteristics (Section 4.3);

As a result of investigating the literature, especially problem-based learning, authentic learning, and automated assessment, the main characteristics of ALAM were proposed by the researcher. These characteristics were discussed with and evaluated by experts. Consequently, the main ALAM characteristics were developed based on the outcome of discussions and formative evaluations of experts. These characteristics include mapping the actions by means of a formal syntax, comparison to multiple expert reference solutions, and feedback for learning.

3. Developing the ALAM components (Section 4.3);

After the objectives were defined and the significance of the solution was studied, ALAM was developed in the 'Design & Development' step of DSR. The ALAM components were based on the main characteristics developed earlier in this step. All the components were developed following consultations with supervisors and experts, and drawing on the researcher's prior knowledge and understanding. The components include the Performance Codifier Engine (PCE), Comparison Engine (CE), and Feedback Compiler Engine (FCE).

4. Developing the conceptual model of an assessment system based on ALAM (Section 4.4).

The concept development of the Action-based Learning Assessment System was based on the researcher's expertise and constant formative feedback and discussions with supervisors and system developers. The first draft was designed by the researcher and discussed with supervisors and experts, followed by the re-design and refinement of the concept. The development process of ALAM is illustrated in Figure 4.1.

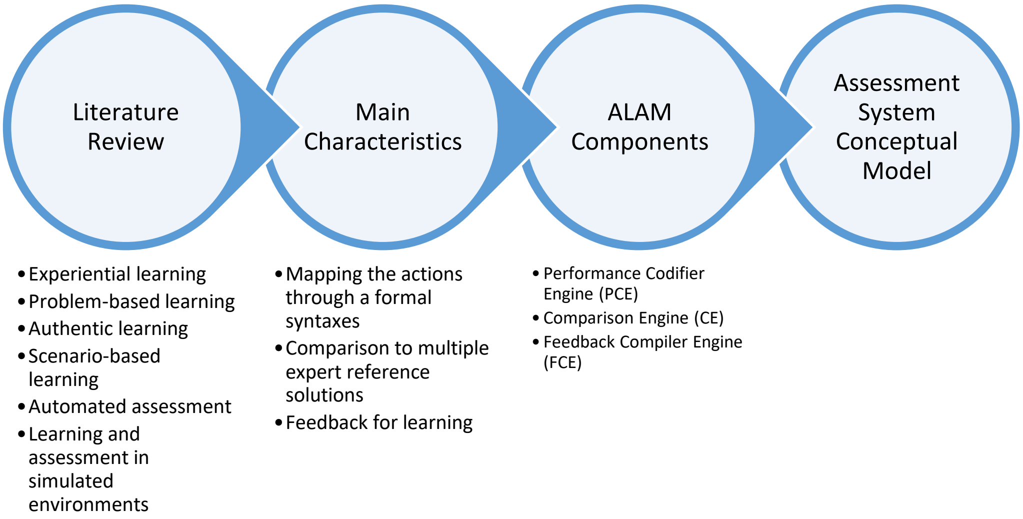

Literature Review
Main Characteristics
ALAM Components
Assessment System Conceptual Model
• Experiential learning
• Problem-based learning
• Authentic learning
• Scenario-based learning
• Automated assessment
• Learning and assessment in simulated environments
• Mapping the actions through a formal syntaxes
• Comparison to multiple expert reference solutions
• Feedback for learning
• Performance Codifier Engine (PCE)
• Comparison Engine (CE)
• Feedback Compiler Engine (FCE)

Figure 4.1: ALAM development process

4.3. The ALAM Main Characteristics and Components

4.3.1. Mapping the Actions through a Formal Syntaxes

ALAM compares the performed actions to reference solutions in order to automatically create a formative feedback for the learner. The performed actions need to have a standard structure so that the learners' and experts' performances are comparable. ALAM uses BEHAVE taxonomy to map the performed actions. The output of the mapping process is a sequence of actions coded with a standard syntax, provided by BEHAVE (Section 5.2.2), that includes the class of action, its type, the action-attributes set to describe the action, dependency rules to show the relation between the actions, and the timestamp to show the sequence of actions.

Structure-mapping theory (Gentner, 1983) suggests a system of "objects, object-attributes and relations between objects" (p. 156) to describe a domain or situation. In the research presented in this thesis, the human actions are the objects referred to in terms of the structure-mapping theory. These three components are included in the BEHAVE syntax to describe performed actions, and ALAM uses this syntax to map the performed actions during the assessment. ALAM uses the syntax to structure the actions as computer-readable data. The defined syntax includes all the required information for the assessment and comparison process (Section 5.2.2). The mapping and codifying process are performed by the PCE.

The outcome of the mapping process is a list of coded actions. Each line of coded action includes the performer (expert or learner) ID, the sequential order of the performed action, and the BEHAVE syntax for that particular action. Sub-section 5.7.1 explains how PCE uses BEHAVE to code the performed actions. This list of sequentially coded actions is used in the comparison process in the next step.

4.3.2. Comparison to Multiple Expert Reference Solutions

To increase the plurality of opinions and procedures, ALAM compares the learner's performance to multiple reference solutions provided by experts in the relevant field. The reference solutions have the same focus of achieving the outcome goal, although they might differ in terms of the procedure used to achieve the goal. This plurality gives learners more freedom to strategise in order to achieve the outcome goal.

A study of various subjects (e.g. decision making, forecasting, project portfolio selection) indicates that accuracy is increased when the number of experts increases (Clemen, 1989; Winkler & Clemen, 2004; Shih et al., 2005; Loh & Sheng, 2014; Roland et al., 2016). The experiments of Winkler & Clemen (2004) suggest that both increases in the number of experts and assessment methods decrease the risk and increase the accuracy of decisions. Winkler & Clemen (2004) reported that "analyses of averages of correlations from multiple experts and/or multiple methods demonstrate a substantial degree of improvement in accuracy as we increase the number of experts or methods" (p. 173). Moreover, Winkler & Clemen (2004) emphasised that "the striking feature […] is the much better performance from multiple experts than from multiple methods" (p. 169). In addition to Winkler & Clemen (2004)'s suggestion to use multiple experts, the use of multiple expert reference solutions was suggested by the participants who undertook the survey for this research (Sections 6.2 and 7.2.1). The survey participants suggested three to five experts. Loh & Sheng (2014) refer to the use of multiple experts as a 'multiple-solution'.

ALAM uses different types of similarity comparisons. The general similarity types are related to the Structure-mapping theory (Gentner, 1983), including:

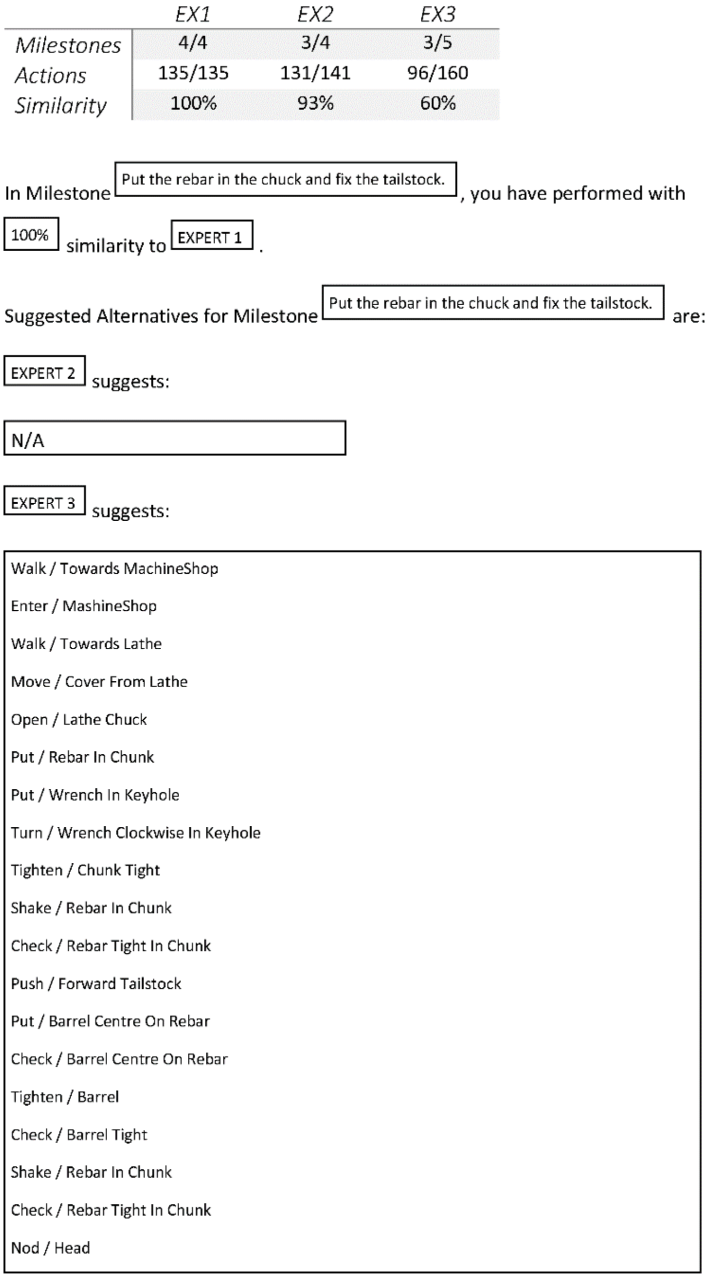
- Literal similarity: the exact mapping of object-attributes and relations;
- Analogy: the mapping of the relations but with few or no object-attributes;
- Abstraction: is similar to analogy but without the object-attributes;
- Anomaly: there no object-attributes or relations to be mapped.

These four general types of similarity concern different actions, their attributes, rules and relations, and action timings and sequences. However, the comparison process is not limited only to the comparison of syntaxes. Each action should be analysed for its relevance. As the performed actions are goal-oriented actions intended to achieve the outcome goal, the relevance of the actions must be studied. The results of the analysis of the actions, done by CE, are used to generate an automated formative feedback for the learners by FCE. Sub-section 5.7.2 explains how CE uses BEHAVE for the comparison process.

4.3.3. Feedback for Learning

ALAM provides a detailed post-performance formative feedback (Figure 4.2) to learners, thereby supporting their performance improvement. ALAM constructs the feedback based on two formative feedback classifications proposed by Rogers (1951) and Shute (2007) introduced in Section 2.3.2. Based on these two classifications, formative feedback can be provided at different levels of complexity due to the amount of information that needs to be given about a learner's performance. The 'Understanding' level of the Rogers (1951) classification of feedback is used by ALAM as a default level of feedback. In this level, ALAM adapts the 'informative tutoring' type of feedback proposed by Shute (2007).

Informative tutoring includes 'verification feedback' (the correctness of response(s), such as right/wrong or overall percentage correct), 'error-flagging' (highlights errors in a solution, without giving the correct answer), and 'strategic hints on how to proceed' (guides the learner in the right direction, yet avoids explicitly presenting the correct answer). However, informative tutoring does not provide the correct answers to the learner. Therefore, ALAM added the correct answer(s) and possible alternative solutions to this type of feedback. The generated feedback allows learners to receive all the information they require to correct and master their practical knowledge.



	EX1	*EX2*	*EX3*
Milestones	4/4	3/4	3/5
Actions	135/135	131/141	96/160
Similarity	100%	93%	60%

In Milestone Put the rebar in the chuck and fix the tailstock., you have performed with

100% similarity to EXPERT 1 .

Suggested Alternatives for Milestone Put the rebar in the chuck and fix the tailstock. are:

EXPERT 2 suggests:

N/A

EXPERT 3 suggests:

Walk / Towards MachineShop
Enter / MashineShop
Walk / Towards Lathe
Move / Cover From Lathe
Open / Lathe Chuck
Put / Rebar In Chunk
Put / Wrench In Keyhole
Turn / Wrench Clockwise In Keyhole
Tighten / Chunk Tight
Shake / Rebar In Chunk
Check / Rebar Tight In Chunk
Push / Forward Tailstock
Put / Barrel Centre On Rebar
Check / Barrel Centre On Rebar
Tighten / Barrel
Check / Barrel Tight
Shake / Rebar In Chunk
Check / Rebar Tight In Chunk
Nod / Head

Figure 4.2: A mock-up of ALAM formative feedback based on example in Section 5.3.2.2

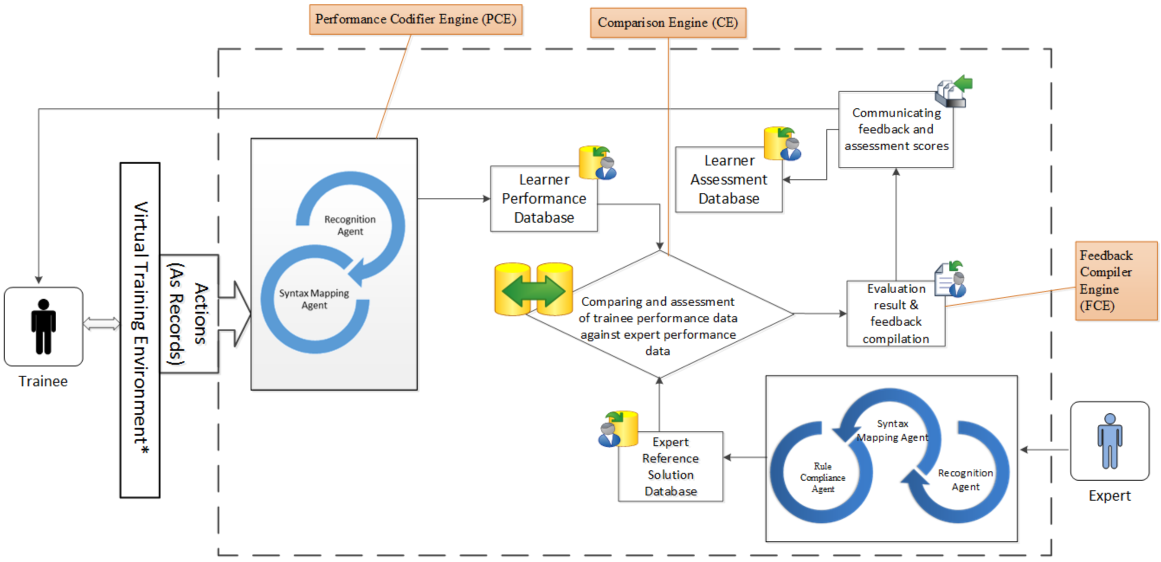
The most straightforward feedback, and the easiest to automatically generate, is a complete (mis)match of the performed actions; however, it is important to include a comment that this is based on the current experts' opinions and does not imply that the new solution could not be right (we assume that the expert solutions only resemble the most likely solutions to solve the problem; other solutions might also be valid; unless specified by restrictive rules).

Of all the possible actions, there is one particular action that is necessary to start a process, and if that action is not performed, it is impossible to initiate the process. A ‘definite failure’ is given if this required rule is not fulfilled; e.g., not turning on the oven for the heating process.

The formative feedback must distinguish between the performance goals and the process goals. If all experts have the same performance goals in their action-sequences, these are required and should be achieved by the learner as well. Thus, the feedback must emphasise such mismatches. The same principle applies to actions that all experts have performed; yet, it is important to note the others that are not done by all experts (e.g., wiping the table after each step) and might not be mandatory. Therefore, there is no need for this action to be done by the learner. Sub-section 5.7.3 explains how FCE uses BEHAVE in the feedback generation process.

4.4. Assessment System Conceptual Model

The conceptual model in Figure 4.3 illustrates a system architecture that combines the three main components of ALAM, the Performance Codifier Engine, Comparison Engine, and Feedback Compiler Engine (explained in Section 4.3), into a workflow starting with the stream of actions and their attributes and ending in the generation of the formative feedback.


Performance Codifier Engine (PCE)
Comparison Engine (CE)
Communicating feedback and assessment scores
Learner Assessment Database
Learner Performance Database
Recognition Agent
Syntax Mapping Agent
Virtual Training Environment*
Actions (As Records)
Trainee
Comparing and assessment of trainee performance data against expert performance data
Evaluation result & feedback compilation
Feedback Compiler Engine (FCE)
Expert Reference Solution Database
Syntax Mapping Agent
Rule Compliance Agent
Recognition Agent
Expert

* Users can use different mediums like Virtual Training Environments, Virtual Worlds, Games, Simulators, and Virtual Reality peripherals.

Figure 4.3: Action-based Learning Assessment System Conceptual Model

Figures 4.4, 4.5, and 4.6 illustrate the wireframe of the Action-based Learning Assessment Systems. As illustrated in Figure 4.3, the actions performed during the assessment can be received in various ways. The actions are tagged with the user information such as role (e.g. expert or learner). The data stream may be fed to the system using input files, external database (e.g. VTE data base), VR peripherals, or in some cases manually.

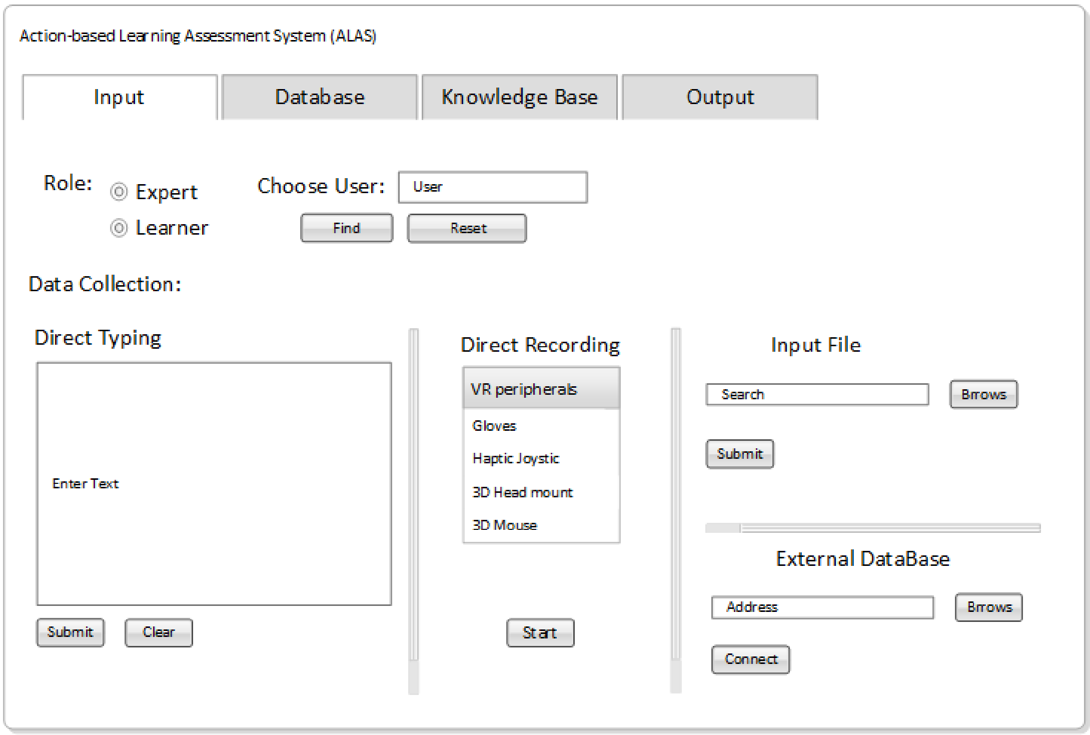


Figure 4.4: Action-based Learning Assessment Systems input interface wireframe

After the data entry has been completed, PCE uses a Recognition Agent (RA) to recognise the actions (class and type), attributes (preposition, adjective, object, quantity, unit, property, location), and timestamps within the input data. Then the Syntax Mapping Agent (SMA) uses the recognised information to create the action syntax. If coding the expert performance, PCE also uses a Rule Compliance Agent (RCE) to analyse the syntax list and manual rules established by the expert to recognise the rules and relations for each action. The results are added to the syntax by SMA. Finally, SMA maps the action-sequences and saves the syntax list on the Learner Performance Database or Expert Reference Solution Database.

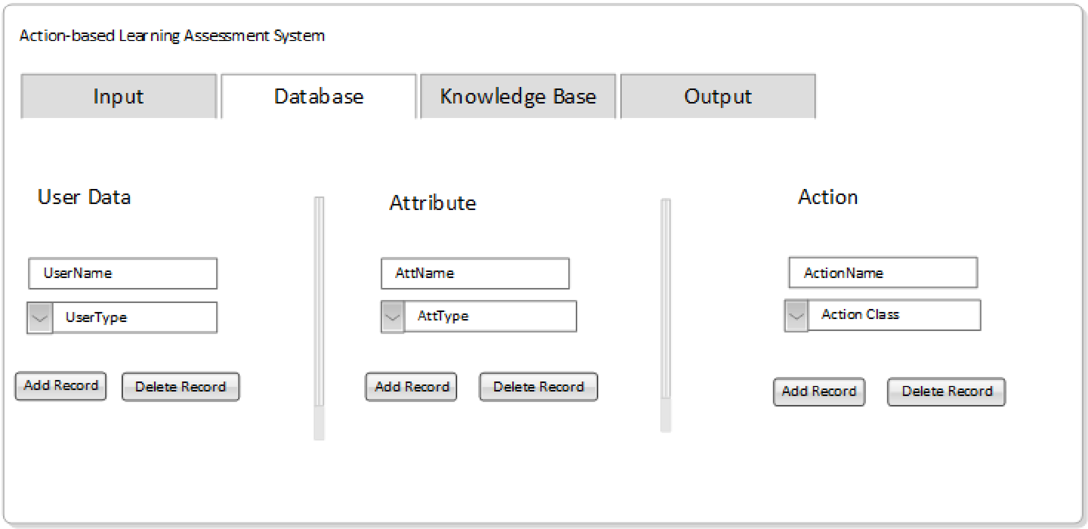


Figure 4.5: Action-based Learning Assessment Systems database interface wireframe

The operator has the option of choosing multiple expert reference solutions with which to compare the learner's performance (Figure 4.5). CE analyses the coded learner actions and stores the analysis results on the database so it can be read by the FCE to compile the formative feedback.

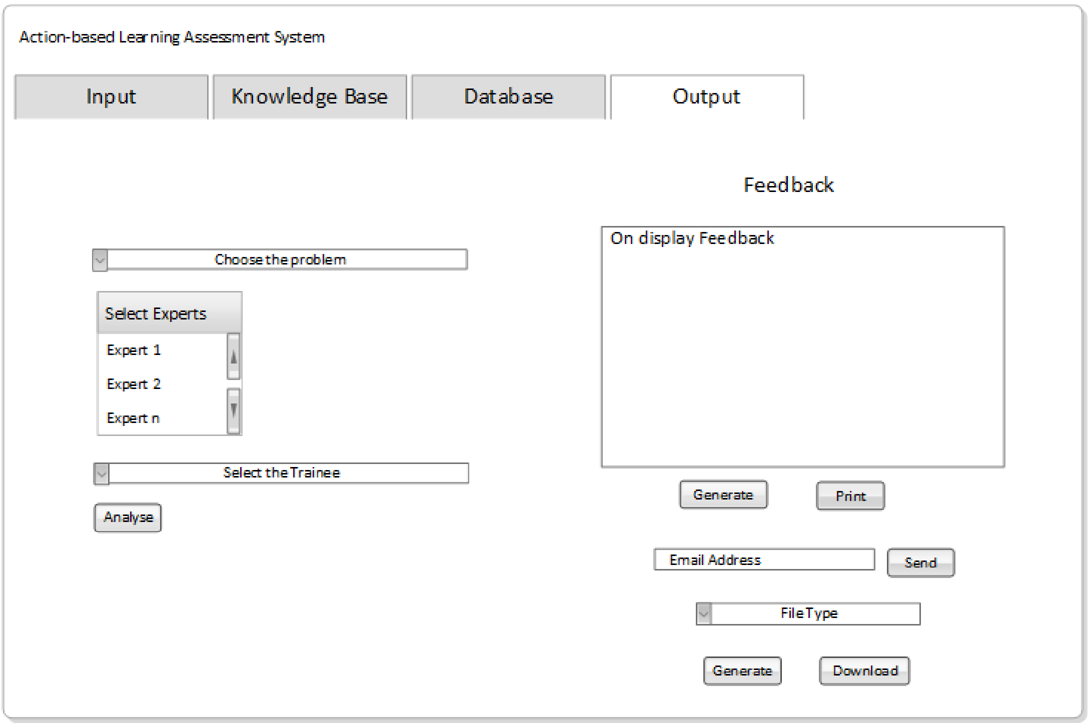


Figure 4.6: Action-based Learning Assessment Systems output interface wireframe

As mentioned in Section 4.3, for the Action-based Learning Assessment System concept to be realised, the first step is the development of BEHAVE. Section 5.7 explains how BEHAVE is used by different components of ALAM.

4.5. Summary

ALAM enables learners to perform in a simulated environment, be compared to multiple reference solutions, and receive a detailed formative feedback that helps them to improve their skills and apply their knowledge to real-life scenarios. This chapter presented the ALAM framework and its conceptual model. The ALAM support for authentic assessment was discussed in this chapter. The concepts of ALAM three components (PCE, CE, and FCE) were presented. However, as the development of PCE, as the starting point of ALAM, depended on the development of BEHAVE, the concepts of CE and FCE were discussed in abstract terms.

In this chapter, the need for a taxonomy of human actions was expressed with the aim of enabling ALAM to code the learners' performed actions, for further analysis and feedback generation. The next chapter introduces BEHAVE taxonomy, its classification, definitions, syntax, and example scenarios. Moreover, the use of BEHAVE in PCE, CE, and FCE is demonstrated.

Chapter 5: BEHAVE: Taxonomy of Human Actions

5.1. Introduction

The assessment of learners' ability to apply their skills in solving problems in practice strongly depends on observing their performed actions, mainly goal-oriented actions. Goal-oriented actions in this context are actions performed in order to accomplish a certain Performance goal.

As discussed in Section 4.3.1, ALAM PCE needs a standard classification of human actions and a formalised syntax to code the performed actions so as to enable CE and FCE to analyse the learner's actions and create a formative feedback. An investigation of the literature in pursuit of finding a suitable taxonomy of human actions (Section 2.5.4) that was appropriate for the researcher's purpose, proved to be futile (Table 2.6 summarises these taxonomies and their shortcomings in terms of the needs of this research). Thereupon, BEHAVE was developed to enable ALAM to codify the learners' actions in the form of computer-readable codes. As illustrated in Figure 3.5, BEHAVE is the artefact developed during the second round of DSR methodology of this research. BEHAVE is the main artefact of this research and was evaluated in the 'Demonstration and Evaluation' step of DSR.

BEHAVE classifies human actions according to various levels and classes and also applies different rules to these actions. These rules mainly present the relations between the actions. BEHAVE uses an action-attributes set to describe how the action has been performed, and a formalised syntax. In this research, given the research scope, BEHAVE focuses more on human actions performed in an Action-based Learning scenario, within a 3D VTE. Nevertheless, BEHAVE is designed to be open and flexible, so that other disciplines can improve upon it or adapt it to their needs and applications.

This chapter explains BEHAVE, beginning with an explanation of the development process, followed by the definition of its levels and classes. Furthermore, a set of action-attributes and a formalised syntax are described. The action-attributes set is employed to describe the performed actions, and the syntax is used to create a standard structured code for each action. Moreover, BEHAVE is applied to different scenarios, including a task analysis scenario, to show its generalisability. There is a further explanation showing how ALAM uses BEHAVE to produce a standard list of coded actions to be used for the generation of automated formative feedback. Furthermore, the effects of context and language on BEHAVE, and taxonomies of human actions in general, are discussed.

5.2. BEHAVE taxonomy

BEHAVE was developed by combining primary and secondary research, collecting data via an expert opinion survey, and investigating the relevant literature. However, the expertise of the researcher played a substantial role in the development of the taxonomy. The secondary research was used as the foundation of BEHAVE, based on the theory of human actions (Goldman, 1970), Taxonomy of Embodied Actions for cooperative design in a distributed company (Robertson, 1997, 2000), Avatar Capabilities Model (Chodos et al., 2014), and other taxonomies that are demonstrated in Table 2.6.

Following an intensive study of the literature and analysis of the previously mentioned theories and taxonomies, the researcher's expertise was used to develop the BEHAVE levels, classes, and the action coding syntax. Direct observation of human actions in both real and simulated environments made a major contribution to the creation of BEHAVE. The primary data collected via the survey was also used for development and evaluation purposes.

BEHAVE classifies human actions according to three main levels and six classes. The levels map the three levels of goals in ALAM (Section 4.2), including the Goal Act, Constitutional Acts, and Functional Acts. The Goal Act consists of one or more Constitutive Acts, which in turn comprise one or more Functional Acts. Functional Acts, as the most basic action level, are classified according to six different classes including Gestural, Responsive, Decisional, Operative, Constructional, and Locomotive.

The following sections explain the levels and classes of the taxonomy, with a subsequent description of the action-attributes set and the syntax. Each of these constituent elements is explained in detail and illustrated with examples.

5.2.1. Definitions

To assess the learners' skills, experts create scenarios representing the outcome goal that needs to be achieved. The achievement of the given outcome goal (The Goal Act) implies the achievement of performance goals (Constitutive Acts) in the correct order by performing appropriate actions, each of which achieves a process goal (Functional Acts). As BEHAVE is intended to be used for different purposes as a taxonomy of human actions, and not just in ALAM, it has its own terminology. The hierarchical levels and classes of BEHAVE are illustrated in Figure 5.1.

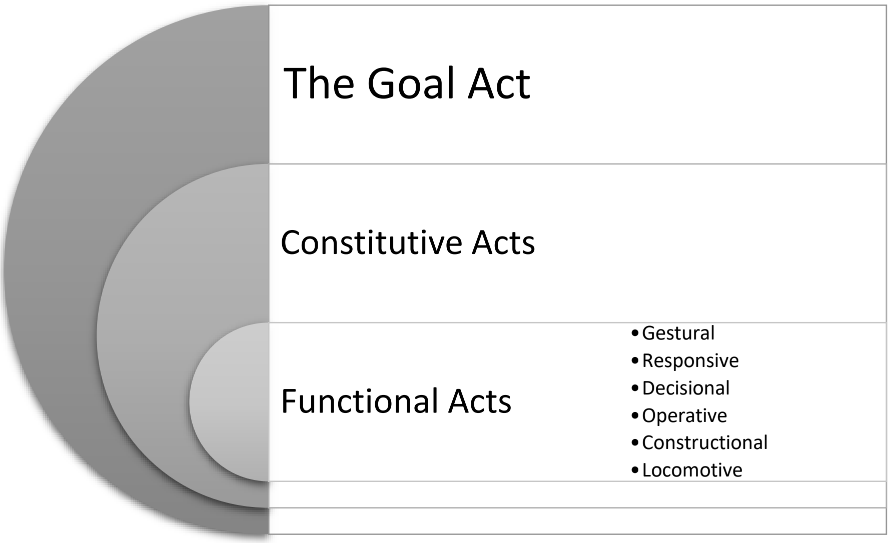


Figure 5.1: The Hierarchical Levels and Classes of BEHAVE

Each hierarchical level of BEHAVE consists of action(s) with the same degree of contribution in the structure of the overall performance. Moreover, each level may include distinctive groups of actions with different overall applications and functions. The following description presents further details on the levels and classes of actions.

The Goal Act

At the highest hierarchical level of actions is the Goal Act that indicates a particular goal to be achieved at the end of the performance. The Goal Act is considered as the highest level of the taxonomy not because of its complexity, but due to its position as the ultimate goal of the performance. The Goal Act (GA) is formed of one or more Constitutive Acts (CA): $GA = \bigcup_{i=1}^{n} CA_i \in CA$

In the context of ALAM, learners must achieve the Goal Act (outcome goal) during the assessment by performing certain actions. These performed actions are mainly aforethought, based on learner's prior knowledge, with the aim of achieving a set of performance goals that ultimately should lead to the achievement of the outcome goal.

Depending on the outcome goal and the assessment scenario, the Goal Act can vary in complexity. The Goal Act can be as complex as “Coronary Artery Bypass Grafting” or as simple as “Put the trash out”. Some of the examples of the Goal Act in the assessment context include: evacuating a library by the fire alarm, stabilising a trauma patient, and disassembling a gun.

Constitutive Acts

To fulfil the Goal Act, a number of simple or compound actions have to be performed and certain objectives achieved. These actions constitute the second level of BEHAVE, namely the Constitutive Acts. The Constitutive Acts consist of atomic actions named Functional Acts (FA): $CA = \bigcup_{i=1}^{n} FA_i \in FA$

Distinguishing the Constitutive Acts that form the Goal Act generally depends on the user’s opinion. However, there should be guidelines to help the users of the taxonomy to create consistency. When analysing the GA to distinguish the CAs, one should bear in mind that CA should not be simplified to the point that it becomes an FA, and it should not be so complex that it becomes the GA. Each CA should have an objective that, along with the other CAs, constitutes the GA.

In the context of ALAM, to achieve the Goal Act (Outcome goal) learners have to break it down into strategically chosen milestones (Performance goals) that their fulfilment leads to the successful Goal Act achievement. These milestones are presented as Constitutive Acts in BEHAVE. Each Constitutive Act realises a certain objective that, together, lead to the Goal Act.

As an example, the Goal Act of ‘evacuating a library by the fire alarm’ consists of ‘form search groups among available staffs on each level’, ‘equip the group members with safety gear’, ‘checking the rooms on each floor’, ‘evacuate non-staff people’, ‘report to the supervisor’, and ‘evacuate the floor as a group’.

Functional Acts

As they are atomic actions, Functional Acts are on the lowest level of the taxonomy. Although BEHAVE considers the Functional Acts as atomic actions, this does not mean that they are as basic as motor skills (e.g. stretching an arm). However, in some cases, a certain Constitutive Act might also be a Functional Act although, in most cases, Constitutive Acts consist of more than one Functional Act.

Functional Acts are classified according to six classes: Gestural, Responsive, Decisional, Operative, Constructional, and Locomotive.

Gestural actions:

Gestures are defined as “a movement that you make with your hands, your head or your face to show a particular meaning” (Gesture, 2016). Gestures are classified as human actions in different taxonomies (Section 2.5.4.3) used for expressing intentions, communicating a meaning, and showing feelings, emotions, or thoughts. These feelings can range from contempt and hostility to approval and affection.

In order to use this class of actions in any application of BEHAVE (e.g. using action recognition in videos surveillance), the meanings of the gestural actions must have a clearly defined context (e.g. culture), as gestures might have different meanings and uses in different contexts. However, regardless of the variations in context-based meaning, these actions are classified as gestural actions.

Responsive actions:

These actions are responses triggered by changes in the environment or by objects. Responsive actions may be involuntary (e.g. reaction to heat, cold, and sharpness) or voluntary (e.g. dodging a coming moving object). Both voluntary and involuntary responses are triggered by stimuli, and the difference is in the process of responding (involuntary actions are reflex actions, while the voluntary actions are more considered); e.g. blench, recede, flinch, recoil, retract, dodge, and wince.

However, Responsive actions should be differentiated from the other classes of actions, such as Gestural actions, in the sense that the purpose of Responsive actions is not to communicate a meaning, or show feelings, emotions, or thoughts.

Decisional actions:

Although behind most performed actions there is some level of decision-making (consciously or unconsciously), BEHAVE recognises a specific class of actions as Decisional actions that enable learners to reflect on their decisions by choosing between options; e.g., between left or right, up or down, yes or no, or on quantities.

Actions such as direct, pick, arrange, check, set, collect, and choose fall under this class of action (Sections 6.3 and 7.2.2).

Operative actions:

Operative actions are simple, atomic actions enabling humans to function and interact with the environment and objects without altering them (e.g. start, carry, and grab).

Constructional actions:

These actions differ from the Operative actions by their manipulation of the objects or the environment (e.g., cut, screw, break, and shatter).

Locomotive actions:

Locomotive actions are used to move around in the environment or to go to different parts of the environment (e.g., walk, run, fly, or teleport).

5.2.2. BEHAVE Taxonomy: Action Attributes and Syntax

A concrete syntax is defined to present the performed actions for both humans and computers. This includes action class, action type, relevant attributes, and possible rules. The syntax provides a formal structure for coding human actions, applying BEHAVE classification of human actions, action-attributes, and defined rules and relationships to the performed actions. The syntax is as follows:

[<Action.Levels>]<Trigger.Action><Action.Class><Action.Type>[Preposition, Adjective, Object, Quantity, Unit, Property, Location][Rules][Timestamp]

Each action is associated with a set of attributes to define the context. Depending on the performed action, any number of relevant attributes would be used. The attributes are discussed in the following:

The Preposition is used to show a relationship in terms of space and time. Some of these prepositions include (but are not limited to): on, in, at, to, by, into, onto, towards, from, of, off, about, through, across, above, over, below, under, beside, until, till, past, before, ago, for, since, next to, out of, and on top of.

(e.g. [*<GA1><CA1><FA1>]<T1><Locomotive><Walk>[Towards*, -, *WaterTap*, -, -, -, -*][09:12:35]*).

The Object is the thing to which the performed action is directed. Although the Object is targeted by the performed action, it might not necessarily be altered.

(e.g. [<GA1><CA1><FA1>]<T1><Locomotive><Walk>[Towards, -, WaterTap, -, -, -, -][09:12:35])

The Adjective is used to qualify the object, Location, or the performed action.

(e.g. [<GA1><CA1><FA1>]<T3><Decisional><Check>[In, Tight, Rebar, -, -, -, Chunk][39SS38][19:16:16])

Property, Quantity, and Unit are used to describe a physical property of the Object or Location, like diameter or temperature.

(e.g. [<GA1><CA1><FA1>]<T2><Constructional><Cut>[-, -, Rebar, 5, cm, Length, Clamp][26SS25][19:10:17])

The Location is used to specify the part of the environment in which the performance is occurring, or where the Object is placed. In some cases (e.g. water), objects in the environment might be considered as Location.

(e.g. [<GA1><CA1><FA1>]<T2><Constructional><Cut>[-, -, Rebar, 5, cm, Length, Clamp][26SS25][19:10:17])

To code the actions executed during a performance, the relations and rules that bound the actions also need to be coded. The Rules used in the BEHAVE syntax are adopted from logical relationships (dependency rules) in project management (Project Management Institute, 2013) including Finish-to-Finish (FF); Finish-to-Start (FS); Start-to-Finish (SF); and Start-to-Start (SS). For example, [26SS25] indicates that action #26 starts at the same time as #25, while [26SF25] means that action #26 starts only if #25 is finished. However, the Rules section of the syntax is open to additional rules, and relations set by the users of BEHAVE depending on their needs. For example, in the case of assessing collaboration in ALAM (Section 4.2.1, Criterion 6) the rules and relations will involve not only the actions, but team members as well. This approach is made possible by the openness and transferability of BEHAVE as a taxonomy.

The Timestamp provides the start time of the action to be performed. The sequential order of the performed actions can be determined based on the Timestamp.

Moreover, the timing of actions can be used to recognise behaviours such as hesitation or uncertainty.

The trigger action identifier at the beginning of each code line of the BEHAVE syntax is generated by the computer, based on the sequence of the performed actions and expert solutions and pre-set rules.

The BEHAVE syntax has been used and tested in an experiment to determine the degree of similarity between the coded scenario and the videotaped and written scenarios (Sections 6.4 and 7.2.3). The majority of participants reported a high degree of similarity with the minimal number of words.

While using the BEHAVE syntax in coding actions performed during the assessment by ALAM, an important point that should be noted is that each element has a specific role in the degree of similarity in comparison with reference solutions (Section 4.3.2). Depending on the degree of similarity required (explained in Section 4.3.2), each element is compared to reference solutions. For example, different levels of similarity use different information:

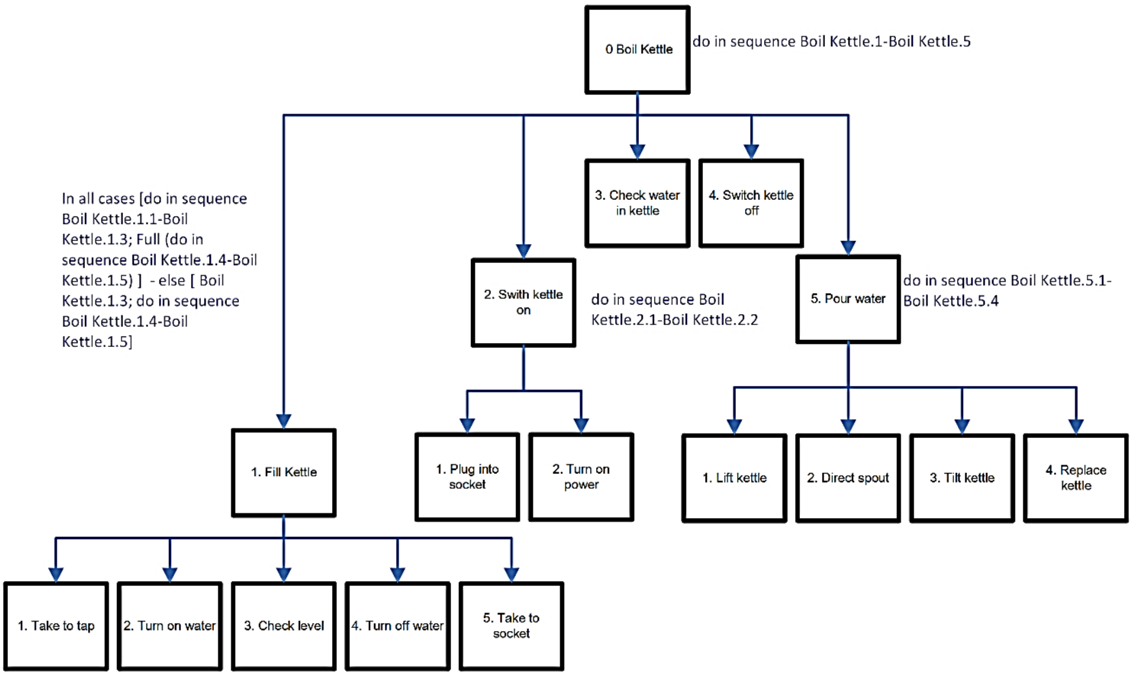
- Literal similarity: *[<Action.Levels>]<Trigger.Action><Action.Class><Action.Type>[Preposition, Adjective, Object, Quantity, Unit, Property, Location][Rules][Timestamp]*;
- Analogy: *[<Action.Levels>]<Trigger.Action><Action.Class><Action.Type>[Preposition, -,Object, -, -, Location][Rules][Timestamp]*;
- Abstraction: *[<Action.Levels>]<Trigger.Action><Action.Class><Action.Type> [Rules][Timestamp]*;
- Anomaly: *[<Action.Levels>]<Trigger.Action><Action.Class><Action.Type>[Timestamp]*.

5.3. Example Scenarios

5.3.1. Applying the Behave Taxonomy to HTA Example

HTA breaks down a task into several operations and sub-operations to "identify those that are likely to fail due to poor design or lack of expertise and thus to propose solutions that might involve redesigning the task or providing special training" (Annett, 2004: p. 33-2). Moreover, HTA does not present any classification of human actions. As Annett (2004) states "HTA analyzes not actions per se but goals and operations, the means of attaining goals" (p. 33-2). However, the structure of an analysed task by HTA including different levels of the operations and their relations may be comparable to the use of hierarchical levels and classes of actions in BEHAVE.

Salmon et al. (2004) created an example for the application of HTA: boiling the kettle. This example is shown in Figure 5.2, and is compared with the same example coded by using BEHAVE.



0 Boil Kettle
do in sequence Boil Kettle.1-Boil Kettle.5
In all cases [do in sequence Boil Kettle.1.1-Boil Kettle.1.3; Full (do in sequence Boil Kettle.1.4-Boil Kettle.1.5)] - else [Boil Kettle.1.3; do in sequence Boil Kettle.1.4-Boil Kettle.1.5]
3. Check water in kettle
4. Switch kettle off
2. Swith kettle on
do in sequence Boil Kettle.2.1-Boil Kettle.2.2
5. Pour water
do in sequence Boil Kettle.5.1-Boil Kettle.5.4
1. Fill Kettle
1. Plug into socket
2. Turn on power
1. Lift kettle
2. Direct spout
3. Tilt kettle
4. Replace kettle
1. Take to tap
2. Turn on water
3. Check level
4. Turn off water
5. Take to socket

Figure 5.2: HTA of the task 'boil kettle' (Created by TaskArchitect3 software)

The Salmon et al. (2004) 'boil kettle' example is mapped to BEHAVE levels and classes of actions as follows:

The Goal Act: Boil Kettle

Constitutive Acts:

1. Fill kettle
2. Boil the water
3. Pour water

Functional Acts:

- Walk towards the water tap (Locomotive)
- Take the kettle to tap water (Operative)
- Turn on water (Operative)

- Check level (Decisional)
- Turn off water (Operative)
- Take to socket (Operative)
- Plug into socket (Operative)
- Turn on the power (Operative)
- Check water in the kettle (Decisional)
- Switch kettle off (Operative)
- Lift kettle (Operative)
- Direct spout (Operative)
- Tilt kettle (Operative)
- Replace kettle (Operative)

As can be seen in this example, the BEHAVE levels and classes are easily mapped to all the different levels of operations and sub-operations in the HTA. The main task is mapped to the Goal Act, the main operations are mapped to the Constitutive Acts, and the sub-operations are mapped to different classes of Functional Acts. The example is coded using the BEHAVE syntax as follows (Table 5.1):

Table 5.1: ‘boil kettle’ example code by BEHAVE syntax

BEHAVE														
The Goal Act	Constitutive Act	Functional Act	Triger Action	Action Class	Action Type	Preposition	Adjective	Object	Quantity	Unit	Property	Location	Rules	Timestamp
1	1	1	1	<Locomotive>	<Walk>	Towards	-	Water Tap	-	-	-	-	[-]	[09:12:35]
1	1	2	-	<Operative>	<Carry>	To	-	Kettle	-	-	-	Water Tap	[2SS1]	[09:12:35]
1	1	3	-	<Operative>	<Put>	Under	-	Kettle	-	-	-	Water Tap	[3SF2]	[09:13:01]
1	1	4	-	<Operative>	<Turn>	-	On	Water Tap	-	-	-	-	[4SF3]	[09:13:08]
1	1	5	-	<Decisional>	<Check>	-	Full	Water Level	-	-	-	Kettle	[5SF4]	[09:13:18]
1	1	6	-	<Operative>	<Turn>	-	Off	Water Tap	-	-	-	-	[6SF5]	[09:13:19]
1	2	7	2	<Locomotive>	<Walk>	To	-	-	-	-	-	Socket	[7SF6]	[09:13:21]
1	2	8	-	<Operative>	<Carry>	To	-	Kettle	-	-	-	Socket	[8SS7]	[09:13:21]
1	2	9	-	<Operative>	<Plug>	Into	-	Kettle	-	-	-	Socket	[9SF8]	[09:13:28]
1	2	10	-	<Operative>	<Switch>	-	On	Power Switch	-	-	-	Kettle	[10SF9]	[09:13:31]
1	2	11	-	<Decisional>	<Check>	In	Boiled	Water	-	-	-	Kettle	[11SF10]	[09:14:46]
1	2	12	-	<Operative>	<Turn>	-	Off	Power Switch	-	-	-	Kettle	[12SF11]	[09:14:47]
1	3	13	3	<Operative>	<Lift>	-	-	Kettle	-	-	-	-	[13SF12]	[09:14:49]
1	3	14	-	<Decisional>	<Direct>	Into	-	Spout	-	-	-	Cup	[14SF13]	[09:14:51]
1	3	15	-	<Operative>	<Steer>	Into	-	Spout	-	-	-	-	[15SF14]	[09:14:52]
1	3	16	-	<Operative>	<Tilt>	-	Downwards	Kettle	-	-	-	-	[16SF15]	[09:14:54]
1	3	17	-	<Decisional>	<Check>	-	Full	Cup	-	-	-	-	[17SF16]	[09:14:57]
1	3	18	-	<Operative>	<Straighten>	-	Upwards	Kettle	-	-	-	-	[18SF17]	[09:14:59]

There are both similarities and differences between BEHAVE and HTA. Although both BEHAVE and HTA include hierarchical levels, the hierarchical levels in HTA do not present any sort of taxonomic levels or classifications. Although HTA breaks down a task into operations and then sub-operations, it does not provide a particular classification. On the other hand, BEHAVE provides both hierarchical levels and a classification of the performed actions.

Both BEHAVE and HTA have sequences. However, HTA uses the sequence of operations regardless of time, while BEHAVE uses timestamps not only for sequencing but to provide information on the time spent on each action.

HTA uses only ‘if…then…’ rule, while BEHAVE benefits from logical relationships (dependency rules) adopted from project management (Section 5.2.2) and is open to rules established by the person who is using the taxonomy.

HTA does not use descriptive attributes for the operations; however, BEHAVE describes each action with a set of attributes to provide a high degree of similarity to the real-life performed action.

Last but not least, although there is HTA software to create HTA diagrams (e.g. TaskArchitect3), their output is limited to a graphical diagram printed on paper or as an image on a computerised device. Unlike BEHAVE, HTA does not generate computer-readable data enabling further computerised analysis (e.g. automated assessment and feedback generation).

5.3.2. Applying the BEHAVE Taxonomy: Other Example Scenarios

In this section, two different scenarios are coded using the BEHAVE syntax to show that the taxonomy can be applied to scenarios from very different fields.

5.3.2.1. Cooking Dinner for the Family (Squash Soup for 4)

The Goal Act is “cooking a squash soup for 4” (i.e., the final state is a prepared and ready-to-serve soup).

The cooking process consists of three *Constitutive Acts*:

1. Preparing for cooking,
2. Preparing the raw ingredients,
3. Mixing and cooking the ingredients.

The following example demonstrates how the BEHAVE taxonomy can be used to code the first Constitutive Act, Preparing for cooking:

Table 5.2: ‘Preparing for cooking’ example code by BEHAVE syntax

BEHAVE														
The Goal Act	**Constitutive Act**	**Functional Act**	**Trigger Action**	**Action Class**	**Action Type**	**Preposition**	**Adjective**	**Object**	**Quantity**	**Unit**	**Property**	**Location**	**Rules**	**Timestamp**
1	1	1	1	<Locomotive>	<Walk>	Into	-	-	-	-	-	Kitchen	[-]	[13:10:21]
1	1	2	-	<Locomotive>	<Walk>	Towards	-	Water Tap	-	-	-	-	[2SF1]	[13:10:23]
1	1	3	-	<Operative>	<Turn>	-	On	Water Tap	-	-	-	-	[3SF2]	[13:10:29]
1	1	4	-	<Operative>	<Put>	Under	-	Hands	-	-	-	Water Tap	[4SF3]	[13:10:31]
1	1	5	-	<Operative>	<Wash>	-	Clean	Hands	-	-	-	-	[5SS4]	[13:10:31]
1	1	6	-	<Decisional>	<Check>	-	Clean	Hands	-	-	-	-	[6SF5]	[13:11:38]
1	1	7	-	<Operative>	<Turn>	-	Off	Water Tap	-	-	-	-	[7SF6]	[13:11:39]
1	1	8	-	<Operative>	<Take>	From	-	Paper	-	-	-	Paper Box	[8SF7]	[13:11:42]
1	1	9	-	<Operative>	<Dry>	-	-	Hands	-	-	-	-	[9SF8]	[13:11:44]
1	1	10	-	<Operative>	<Put>	Into	-	Paper	-	-	-	Bin	[10SF9]	[13:12:09]
1	1	11	-	<Locomotive>	<Walk>	Towards	-	Table	-	-	-	-	[11SF10]	[13:12:12]
1	1	12	-	<Operative>	<PickUp>	From	-	Recipe	-	-	-	Table	[12SF11]	[13:12:15]
1	1	13	-	<Operative>	<Read>	-	-	Recipe	-	-	-	-	[13SF12]	[13:12:17]
1	1	14	-	<Operative>	<Put>	On	-	Recipe	-	-	-	Table	[14SF13]	[13:13:19]

The coded Functional Acts above are performed by one of the experts to achieve a Constitutive Act: preparing for cooking. In the context of ALAM, one has to consider that this list of coded actions might differ from expert to expert in terms of sequence or the particular actions chosen. As long as the learner achieves the CAs and consequently the GA, s/he receives a formative feedback after the performance has been compared with expert reference solutions.

5.3.2.2. Supporting Copper Rebar in the Lathe Chuck

The Goal Act is “Supporting copper rebar in the lathe chuck” and the trainee needs to perform three *Constitutive Acts* successfully. To support the rebar in the lathe machine, the trainee has to:

1. Enter the shop and collect safety equipment and clothing;
2. Choose and size the cooper rebar;
3. Put the rebar in the chuck and fix the tailstock.

The following example demonstrates how the BEHAVE taxonomy can be used to code Constitutive Act 3, Put the rebar in the chuck and fix the tailstock:

Table 5.3: ‘Put the rebar in the chuck and fix the tailstock’ example code by BEHAVE syntax

BEHAVE														
The Goal Act	Constitutive Act	Functional Act	Trigger Action	Action Class	Action Type	Preposition	Adjective	Object	Quantity	Unit	Property	Location	Rules	Timestamp
1	1	29	3	<Locomotive>	<Move>	Towards	-	-	-	-	-	Machine Shop	[29SF28]	[19:13:45]
1	1	30	-	<Operative>	<Enter>	-	-	-	-	-	-	Machine Shop	[30SF29]	[19:13:56]
1	1	31	-	<Locomotive>	<Move>	Towards	-	Lathe	-	-	-	-	[31SF30]	[19:14:00]
1	1	32	-	<Operative>	< Move>	From	-	Cover	-	-	-	Lathe	[32SF31]	[19:14:13]
1	1	33	-	<Operative>	<Open>	-	-	Chuck	-	-	-	Lathe	[33SF32]	[19:14:56]
1	1	34	-	<Operative>	<Put>	In	-	Rebar	-	-	-	Chunk	[34SF33]	[19:15:18]
1	3	35	-	<Operative>	<Put>	In	-	Wrench	-	-	-	Key Hole	[35SF34]	[19:15:26]
1	3	36	-	<Operative>	<Turn>	-	Clockwise	Wrench	-	-	-	-	[3SF35]	[19:15:36]
1	3	37	-	<Operative>	<Tighten>	-	Tight	Chunk	-	-	-	-	[37SS36]	[19:15:36]
1	3	38	-	<Operative>	<Shake>	In	-	Rebar	-	-	-	Chunk	[38SF37]	[19:16:16]
1	3	39	-	<Decisional>	<Check>	In	Fixed	Rebar	-	-	-	Chunk	[39SS38]	[19:16:16]
1	3	40	-	<Operative>	<Push>	Forward	-	Tailstock	-	-	-	-	[40SF39]	[19:16:22]
1	3	41	-	<Operative>	<Put>	On	Centred	Barrel	-	-	-	Rebar	[41SF40]	[19:16:28]
1	3	42	-	<Decisional>	<Check>	On	Centred	Barrel	-	-	-	Rebar	[42SS41]	[19:16:29]
1	3	43	-	<Operative>	<Tighten>	-	Tight	Barrel	-	-	-	-	[43SF42]	[19:16:45]
1	3	44	-	<Decisional>	<Check>	-	Tight	Barrel	-	-	-	-	[44SF43]	[19:17:15]
1	3	45	-	<Operative>	<Shake>	In	-	Rebar	-	-	-	Chunk	[45SF44]	[19:17:20]
1	3	46	-	<Decisional>	<Check>	In	Tight	Rebar	-	-	-	Chunk	[46SS45]	[19:17:20]
1	3	47	-	<Gestural>	<Nod>	-	-	Head	-	-	-	-	[47SF4]	[19:17:25]

The examples given above demonstrate the versatility of the performances that can be coded by BEHAVE.

5.4. Actions in Different Contexts

As discussed in Section 2.5.2, actions may have various meanings depending on their contexts. As Salmon et al. (2008) stated "human action occurs as an event within a context. By understanding the relationship between these three entities (i.e. the relationships between human action, device states and context of the interaction), one might begin to propose a contextual theory of human action" (p. 18).

Salmon et al. (2008) categorised the context as internal or external. "Internal context (cf. knowledge in the head) refers to the knowledge, beliefs, experiences and motivations of the individual concerned. External context (cf. knowledge in the world) refers to the situational, temporal, informational, design and environmental characteristics present" (p. 18).

Mackieh & Cilingir (1998) and Kim & Jung (2003) highlighted the role of the internal and external contexts in Human Error Analysis (HEA) or Human Reliability Analysis (HRA) in safety assessment. Actions and activities performed by humans are usually influenced by "given specific working conditions or task situations, so-called context, which is comprised of the MTO (man, technology and organization) triad" (Kim & Jung, 2003: p. 479). They introduced different terminologies for conditions that affect human performance. These influencing conditions have been represented via numerous 'context factors' according to the method being used, including "PSF (performance shaping factors), PIF (performance influencing factors), IF (influencing factors), PAF (performance affecting factors), EPC (error producing conditions), CPC (common performance conditions), and so on" (Kim & Jung, 2003: p. 479).

Although the Human Error Identification (HEI) research literature is more involved with the identification and classification of mistakes and errors and not the actions per se, the context of the performance that leads to an error is critical and a well-recognised factor. Actions and their failure to achieve a goal do not make much sense without knowing the context in which they occurred, "and since the context often may be the 'error forcing condition' that leads to the failure, it seems reasonable to consider how the coveted 'error probability' can be determined directly from a characterisation of the context." (Fujita & Hollnagel, 2004: p. 146).

Actions might convey different meanings in different disciplines, e.g. the action run is locomotive (moving at a speed faster than a walk, never having both feet on the ground at the same time) in a running competition or in response to something life threatening. ‘Run’ may deliver the meaning of running a machine as an operation. It could be used in the context of politics such as running for the presidency. The meaning of the verb conveying the action depends upon the context in which the action occurs. The significance of different contexts is discussed in the following section.

5.5. (English) Language Effects on the Taxonomy of Human Actions

Human action theories and taxonomies are limited by language, especially verbs and nouns (e.g. run, set, pick), with respect to the exhaustiveness of human actions and performances (Reed, 1967; Oller, 1968; Goldman, 1970; Bennett, 1971; Fleishman et al., 1984).

Table5.4: Different meanings of an action leading to different classifications – results from the card sorting test explained in Section 6.4 (smaller percentages are not shown in this table, so they will not add up to 100%)

	Constructional	Gestural	Locomotive	Operative	Decisional	Responsive
Drop	52%	-	-	27%	-	12%
Drag	50%	-	-	35%	-	-
Throw	47%	-	-	33%	-	-
Drive	-	-	77%	17%	-	-
Jump	-	-	64%	13%	-	21%
Move	-	-	28%	63%	-	-
Talk	-	14%	-	61%	-	12%
Turn	-	-	16%	55%	19%	-
Pick	-	-	-	16%	77%	-
Direct	-	-	-	14%	71%	-
Arrange	14%	-	-	14%	67%	-
Collect	-		-	28%	58%	-
Set	11%	-	-	22%	57%	-
Wince	-	21%	-	-	-	76%

BEHAVE also faces the same language constraints (Table 5.4). There are various possible approaches to deal with this problem including but not limited to:

Pre-defined action dataset: Creating a specific dataset and classifying all the possible actions and using this dataset as a reference for classifying actions under different functional classes which limit the use of taxonomy in different contexts.

Automated essay grading vocabulary databases: the current vocabulary databases from various automated essay grading technologies might be useful and may contribute to the decrease of the language effect on the taxonomy.

Functional Acts class expansion: expanding the classes to more categories, thereby creating the major problem of having very similar classes with slight differences, which makes the classification process far more complicated and reduces the practicality of the taxonomy.

Specialised datasets: currently, the more practical solution to this issue appears to be a standard data set that delivers the context flexibility for such actions through the use of synonyms. Lexical databases such as WordNet can be used to find appropriate synonyms for each action that has multiple contextual meanings. In the context of ALAM and other computerised applications of BEHAVE with specific contexts, the use of specialised data sets may be the better choice.

Omission of the language: as the BEHAVE syntax is mainly intended to create computer-readable codes for each action, one solution might be to omit the language and map the concepts to the computer generated codes. For example, if the person grabs an object, stretches the arm backwards, and moves the arm forward and releases the object, this would be the coded concept T1O6, which means the action ‘throw’.

Experience API[1] Statement Properties: as Wild et al. (2014) states, "Each ‘action’ has a ‘predicate’, which is the verb required for inserting trace statements to the xAPI [(“Experience API, Version 1.0.1,” 2013)] tracking endpoint" (p.27). Experience API uses various statement properties including verb and context (“Experience API, Version 1.0.1,” 2013). xAPI uses “Internationalized Resource Identifier (IRI)” to correspond to the meaning of the verb, and not the word. Verbs are consisted of “an IRI and a set of display names corresponding to the multiple languages or dialects which provide human–readable meanings of the Verb" (“Experience API, Version 1.0.1,” 2013: p.14). The IRI represents a particular semantic of a word and not the word. Each statement in xAPI has an optional field to

[1] "The Experience API is a service that allows for statements of experience to be delivered to and stored securely in a Learning Record Store (LRS). These statements of experience are typically learning experiences, but the API can address statements of any kind of experience. The Experience API is dependent on Activity Providers to create and track these learning experiences; this specification provides a data model and associated components on how to accomplish these tasks" (“Experience API, Version 1.0.1,” 2013: p.3).

provide contextual information. The use of ‘Experience API’ statement properties (verb, semantics, and context) might assist BEHAVE to overcome the barrier of language and context.

5.6. Need and Use of the BEHAVE Taxonomy

In the initial stage of this research, the possibility of assessing the learners using automated assessment based on their performed actions instead of a written essay was raised. This question led to the development of a new assessment method named Action-based Learning Assessment Methodology (ALAM), explained in Chapter 4.

During the development of ALAM, there emerged the need for a standard classification of actions for the purpose of coding the performed actions, regardless of the source of the input data of these actions. The solution was the development of a taxonomy of human actions. The first step to developing such taxonomy was an extensive literature review. As discussed in Sections 2.5 and 2.6, most of the taxonomies in the literature were too general (e.g. tasks and performances) or too focused and field-specific (e.g. daily actions in the kitchen or gestures used in touch screens); therefore, they could not be adopted (See Table 2.6). However, the current taxonomies and classifications have contributed to the development of BEHAVE.

In addition to the primary necessity for the taxonomy (ALAM), the literature review opened up an opportunity to contribute to other fields such as error recognition, task analysis, video tagging and action recognition in videos. These fields lacked such taxonomy; hence, the necessity to develop an exhaustive taxonomy to classify human actions. One might argue that each research or field can benefit from the development of a focused field-specific taxonomy as this is common practice among researchers. However, the existence of several partial taxonomies in each field may result in confusion and failure of communication. Therefore, the development of BEHAVE can help to overcome these issues.

BEHAVE is designed to be flexible, transferable, and exhaustive. Furthermore, the taxonomy covers not only the performed actions in 3D VTE; real-life actions can be classified by this taxonomy. Obviously, the taxonomy is within the boundaries of the current use and present research. However, in Section 8.4 of the concluding chapter, suggestions are offered for future research and development directions.

5.7. BEHAVE Application in ALAM

ALAM delivers an automated formative feedback using multiple experts' opinions based on learners' performances in 3D VTE. As an automated assessment methodology, ALAM needs a computerised platform to function at its best. (2D/3D) VTEs can, directly or by means of third party software, provide actions as digital entities which can be coded by the Performance Codifier Engine, and processed further by the Comparison Engine leading to an automated formative feedback, generated by the Feedback Compiler Engine.

To be able to compare learners' performances with the experts' reference solutions, ALAM requires a list of actions, their attributes, and action-sequences that constituted the performance. Furthermore, this information has to be in the form of computer-readable codes in order to be usable by ALAM's different engines.

BEHAVE provides a precise classification of performed actions and a formalised syntax that enables ALAM to code the performed actions that can then be analysed by a computerised system. ALAM uses different classes of actions and the action-attributes set to create a precise description of an action with similar details of the actions performed by the learner in the (2D/3D) VTE as the representation of the real-life actions. The BEHAVE syntax codes the performed actions in detail, restrains them with rules and relations, and records their temporal sequences by the use of timestamps. ALAM benefits from this formalised representation of performed actions of both learners and experts that makes the comparison feasible and consistent.

5.7.1. Performance Codifier Engine

The PCE component receives a stream of data from a source (e.g., VR peripherals and (2D/3D) VTE). Whilst checking the input data stream, PCE uses the data sets created based on BEHAVE to detect the actions according to the BEHAVE taxonomy, the relevant attributes (Preposition, Adjective, Object, Quantity, Unit, Property, Location), and the timestamps of the performed action. The sequence of actions is based on the timestamps.

The validity of the sequences is verified by a set of rules that are either manually specified by the experts or deducted automatically from multiple 'approved action-sequences'. The system shown in Figure 4.3 may include an artificial intelligent agent that derives a rule about predecessor relations by analysing the occurrence of a sequence of actions in a particular order (as part of the expert action-sequences). Additional rules and relations may be stated for a certain action-sequence that is used only when the sequence is more similar to an alternative solution. Rules and relations are stated on different levels to match Constitutive and Functional Acts. For example, the learner cannot cook the soup before chopping and adding the raw ingredients. Therefore, all approved action-sequences set by the experts should specify the inclusion of these actions and their appropriate order. The rules and relations are used by FCE to generate the formative feedback (e.g. 'cooking' cannot be performed before 'chopping' the 'raw ingredients').

Constitutive Acts must be defined by the experts as different milestones or in the context of ALAM, as performance goals. The detection of Constitutive Acts would be done by trigger actions[1]. For example, the *<Decisional><Check><-, Boiling, Water, -, -, -, Pot>* defines the end of the Constitutive Act of heating the water; experts define the trigger actions when creating the assessment scenario. The PCE labels these action-sequences as Constitutive Acts to be used by the CE for analysis. Table 5.5 maps the ALAM three primary goals to BEHAVE levels in a scenario coded by BEHAVE syntax.

[1] Trigger action is an action that happens during the performance that is necessary to start or/and end a Constitutive Act.

Table 5.5: ‘Boil Kettle’ ALAM assessment scenario code by BEHAVE syntax

ALAM			BEHAVE														
Outcome Goal	**Performance Goal**	**Process Goal**	**The Goal Act**	**Constitutive Act**	**Functional Act**	**Trigger Action**	**Action Class**	**Action Type**	**Preposition**	**Adjective**	**Object**	**Quantity**	**Unit**	**Property**	**Location**	**Rules**	**Timestamp**
Boil Kettle	Fill kettle	Walk towards the water tap	1	1	1	1	<Locomotive>	<Walk>	Towards	-	Water Tap	-	-	-	-	[-]	[09:12:35]
Boil Kettle	Fill kettle	Take the kettle to tap water	1	1	2	-	<Operative>	<Carry>	To	-	Kettle	-	-	-	Water Tap	[2SS1]	[09:12:35]
Boil Kettle	Fill kettle	Put the kettle under the tap	1	1	3	-	<Operative>	<Put>	Under	-	Kettle	-	-	-	Water Tap	[3SF2]	[09:13:01]
Boil Kettle	Fill kettle	Turn on water	1	1	4	-	<Operative>	<Turn>	-	On	Water Tap	-	-	-	-	[4SF3]	[09:13:08]
Boil Kettle	Fill kettle	Check the water level to be full	1	1	5	-	<Decisional>	<Check>	-	Full	Water Level	-	-	-	Kettle	[5SF4]	[09:13:18]
Boil Kettle	Fill kettle	Turn off water	1	1	6	-	<Operative>	<Turn>	-	Off	Water Tap	-	-	-	-	[6SF5]	[09:13:19]
Boil Kettle	Boil water	Walk to the electricity socket	1	2	7	2	<Locomotive>	<Walk>	To	-	-	-	-	-	Socket	[7SF6]	[09:13:21]
Boil Kettle	Boil water	Carry the kettle to the electricity socket	1	2	8	-	<Operative>	<Carry>	To	-	Kettle	-	-	-	Socket	[8SS7]	[09:13:21]
Boil Kettle	Boil water	Plug in the kettle	1	2	9	-	<Operative>	<Plug>	Into	-	Kettle	-	-	-	Socket	[9SF8]	[09:13:28]
Boil Kettle	Boil water	Turn on the kettle	1	2	10	-	<Operative>	<Switch>	-	On	Power Switch	-	-	-	Kettle	[10SF9]	[09:13:31]

Boil Kettle	Boil water	Check if the water has boiled	1	2	11	-	<Decisional>	<Check>	In	Boiled	Water	-	-	-	Kettle	[11SF10]	[09:14:46]
Boil Kettle	Boil water	Turn off the kettle	1	2	12	-	<Operative>	<Turn>	-	Off	Power Switch	-	-	-	Kettle	[12SF11]	[09:14:47]
Boil Kettle	Pour water	Take the kettle	1	3	13	3	<Operative>	<Lift>	-	-	Kettle	-	-	-	-	[13SF12]	[09:14:49]
Boil Kettle	Pour water	Direct the spout into the cup	1	3	14	-	<Decisional>	<Direct>	Into	-	Spout	-	-	-	Cup	[14SF13]	[09:14:51]
Boil Kettle	Pour water	Steer the spout into the cup	1	3	15	-	<Operative>	<Steer>	Into	-	Spout	-	-	-	Cup	[15SF14]	[09:14:52]
Boil Kettle	Pour water	Tilt the kettle downwards	1	3	16	-	<Operative>	<Tilt>	-	Downwards	Kettle	-	-	-	-	[16SF15]	[09:14:54]
Boil Kettle	Pour water	Check if the cup is full	1	3	17	-	<Decisional>	<Check>	-	Full	Cup	-	-	-	-	[17SF16]	[09:14:57]
Boil Kettle	Pour water	Replace kettle	1	3	18	-	<Operative>	<Straighten>	-	Upwards	Kettle	-	-	-	-	[18SF17]	[09:14:59]

5.7.2. Comparison Engine

The CE compares the performed action sequences of the learner with the solutions offered by multiple expert. A comparison between all Constitutive Acts and Functional Acts is used to calculate the similarity of two sequences. The similarity calculation includes factors with an impact on the final score. Furthermore, low partial similarities of sequences are recorded for the generation of the formative feedback, including the sequences and their attributes. Specifically, the Comparison Engine analyses the coded actions to check for:

- **Non-compliance of rules**: Rules are either strict (i.e., all experts have the same sequence of actions) or loose (i.e., only some experts have the same sequence of actions). Strict rules must be followed; loose rules represent alternatives, such that an exact match is not required.
- **Attributes that do not match**; including a weighting of the relevance of an attribute
- **The timing of the actions**; i.e., the length of time between two actions indicates that there are problems in deciding what to do next.
- **The sequence of Constitutional Acts in comparison to the experts**.
- **Achievement of the Goal Act**.

Although an unachieved Goal Act is considered as a failure, the failure may be caused by an unfulfilled Constructive Act, even if the rest of the Constructive Acts are correct. In this case, the feedback should note that there was a demonstrated high degree of similarity for most actions, and the failure can be avoided by rectifying the wrong Constructive Act.

The coded actions and Constitutive Acts created by PCE, using the BEHAVE syntax, are treated as a string (e.g. {1,2,3,4,5,6,7,8}) in CE, that is compared with multiple strings from the experts. ALAM suggests the use of the Maximum Similarity Index (MSI) by Loh & Sheng (2014) for both Constitutive and Functional Acts. MSI uses the Jaccard Index or JACC (Jaccard, 1912) to compare the action-sequences. JACC first converts the sequences to bigrams, then, divides the size of the sets' intersection by the size of their union: JACC (A, B) = $\frac{|A \cap B|}{|A \cup B|}$. The JACC's value varies between 0 to 1, with 0 being the complete dissimilarity and 1 the complete similarity. MSI uses the greatest JACC (JACCmax) as the most similar sequence.

For example, if the learner performs the Constitutive Acts in the order of 1→2→3→4→5, the sequence is A={12345} and the bigrams are {12,23,34,45}. The results of the comparison with the experts' solutions E1={123465}={12,23,34,46,65} and E2={123456}={12,23,34,45,56} are:

JACC (A,E1) = $\frac{|12,23,34|}{|12,23,34,45,46,65|} = \frac{3}{6} = 50\%$

JACC (A,E2) = $\frac{|12,23,34,45|}{|12,23,34,45,56|} = \frac{4}{5} = 80\%$

As can be seen from the results, the learner's sequence is close to that of expert 2. However, the sequence is partially identical right up to the Constructive Act 6. This learner might be, as Loh & Sheng (2014) suggest, 'likely-expert'. In cases of partial similarity, there is a chance that the learner is likely-expert or has the highest similarity (e.g. A={12345}, E3={12345671} so there are two possibilities: {12345671} which is a 100% match and {12345761} that is a 40% match). MSI deals with the partial sequences by using the Adjusted JACCmax that is the lower possibility.

5.7.3. Feedback Compiler Engine

FCE generally uses the analysis results from the CE to create the feedback without the direct use of BEHAVE. However, FCE may use an artificial intelligence agent to convert the BEHAVE syntax into a sentence-like structure to communicate the correct or alternative solutions to the learner. For example, the code *<Decisional><Check>[In, Tight, Rebar, -, -, -, Chunk]* is converted to Check/Rebar Tight In Chunk.

5.8. Summary

In this chapter, BEHAVE, the taxonomy of human actions that was developed in this research was explained in detail. The levels and classes of the taxonomy were defined, and the taxonomy was applied to different examples. Furthermore, a formalised syntax, its structure, and purpose were discussed. Various issues involving taxonomies of human actions such as context and language were discussed as well. The chapter then explained the need for BEHAVE and how ALAM uses the BEHAVE components, PCE, CE, and FCE.

In the next two chapters (Chapters 6 and 7), the results of the evaluation of BEHAVE are presented, analysed, and discussed. Chapter 6 presents the statistical results and analysis of the evaluation results, and Chapter 7 discusses those results.

Chapter 6: Data and Analysis

6.1. Introduction

As discussed in Sections 2.5.5 and 3.4.3, the evaluation of the taxonomy and artefact are crucial phases in both taxonomy development and DSR methodology. In this research, the BEHAVE taxonomy was evaluated as the DSR artefact and the data was gathered by means of various evaluation methods and tests including an experts' opinion survey, card sorting test, performance coding experiment, and participant feedback. This chapter summarises and analyses the results of the evaluations using descriptive statistics such as tables and graphs, followed by discussions in Chapter 7.

The following inferential statistical techniques were used (Section 3.6 explains the justification of each technique):

1. Survey: Cronbach's Alpha reliability coefficient; Friedman's test for prioritising; Chi-square goodness of fit; Binomial test; Chi-Square Test of Homogeneity;
2. Card Sorting: Scatterplot matrix; K-mean Cluster analysis; Pearson correlation test; Fowlkes and Mallows index; R-squared;
3. Performance Coding Experiment: One-sample t-test; two independent sample t-test; Kolmogorov-Smirnov normality test.

The statistical results presented in this chapter were processed by SPSS and SynCaps V.3.

6.2. Survey Results

A survey on the importance of BEHAVE, its levels and classes, and ALAM elements, was sent to two groups (G) of experts[1] (See Section 3.7 for method of choosing the populations and sampling). One group consisted of virtual worlds experts from the VWWG group in Australia and New Zealand. A second group of respondents were experts from different industries around the world who were invited to contribute by email the perspective of various companies, factories, and universities.

[1] All the responded experts have indicated a relevant field according to the group they were in.

In the following sub-sections, the results of the survey are presented, and the results for the two groups are compared in order to study any differences of opinion (VW and industry). Due to the hierarchical nature of the taxonomy levels and the different intended uses of each level, the results for the questions (Q1-3) addressing these levels are shown separately from the rest of the survey questions.

6.2.1. Survey Results for All the Respondents

The survey consisted of 16 questions. The Cronbach’s Alpha reliability coefficient for all the respondents is equal to 0.888 which satisfies the criterion of the reliability coefficient having to be higher than 0.7 (Cronbach, 1951) for acceptable survey reliability.

6.2.1.1. Importance of Each Level of the Taxonomy

For the first three questions of the survey, asking how important it is to have the Goal Act, Constitutive Acts, and Functional Acts in any Action-based Learning scenario, the Chi-square goodness of fit test is used to investigate whether the hypotheses fit the observed data. The hypotheses are as follows:

Null hypothesis (H0): Respondents have no tendency towards the importance of the subject of Qi.

Alternative hypothesis (H1): Respondents have a tendency towards the importance of the subject of Qi.

i = 1,…, 3;

The expected frequency of the chosen options by the respondents is equal to the number of respondents divided by the number of the chosen options (Table 6.1, Table 6.2, and Table 6.3). This is because if there is no preference, then the probability of each option is the same, and the number of respondents who chose each option should be equal. On the contrary, the observed frequencies show apparent differences in choices among the respondents.

Table 6.1: Chi-square test descriptive statistics for the Goal Act

How important is it to have a clear goal to achieve, in any 'Action-based Learning' scenario?*			
	Observed N	Expected N	Residual
somewhat important	5	12.0	-7.0
Important	6	12.0	-6.0
Very important	25	12.0	13.0
Total	36		

Table 6.2: Chi-square test descriptive statistics for the Constitutive Acts

How important is it to break down the solution into different Constitutive Acts?*			
	Observed N	Expected N	Residual
very unimportant	1	7.2	-6.2
Unimportant	1	7.2	-6.2
somewhat important	8	7.2	.8
Important	12	7.2	4.8
very important	14	7.2	6.8
Total	36		

Table 6.3: Chi-square test descriptive statistics for the Functional Acts

How important is it to break down the milestones into Functional Acts?*			
	Observed N	Expected N	Residual
very unimportant	1	7.2	-6.2
Unimportant	3	7.2	-4.2
somewhat important	9	7.2	1.8
Important	11	7.2	3.8
very important	12	7.2	4.8
Total	36		

Table 6.4, Table 6.5, and Table 6.6 show the Chi-square test outputs including Chi-square statistic and Asymptotic Significance. For all the three questions, the Asymptotic Significance is less than the 0.05 significance level, so the null hypothesis is rejected, and the alternative is accepted. Respondents appeared to have affirmed the importance of the subject of Qi.

Table 6.4: Chi-square test statistics and results for the Goal Act

Test Statistics	
Chi-Square	21.167a
df	2
Asymp. Sig.	.000

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 12.0.

Table 6.5: Chi-square test statistics and results for the Constitutive Acts

Test Statistics	
Chi-Square	20.389a
df	4
Asymp. Sig.	.000

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 7.2.

Table 6.6: Chi-square test statistics and results for the Functional Acts

Test Statistics	
Chi-Square	13.444a
df	4
Asymp. Sig.	.009

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 7.2.

Therefore, chi-square test confirms the differences seen in the descriptive statistics table and shows that:

A high percentage of respondents (about 69%) chose ‘very important’ for having the Goal Act in Action-based Learning scenarios, and 17% of respondents chose’ important’ (Figure 6.1).

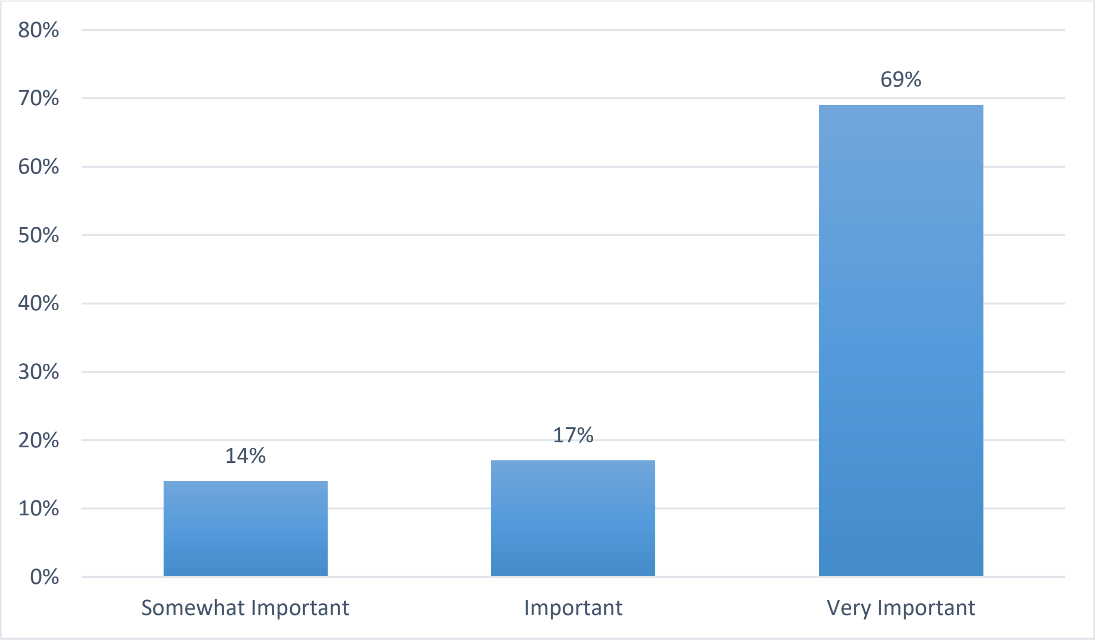


Figure 6.1: Respondents’ views about the importance of the Goal Act

A high percentage of respondents (about 39%) chose ‘very important’ for having Constitutive Acts in Action-based Learning scenarios, and 33% of respondents chose ‘important’ (Figure 6.2).

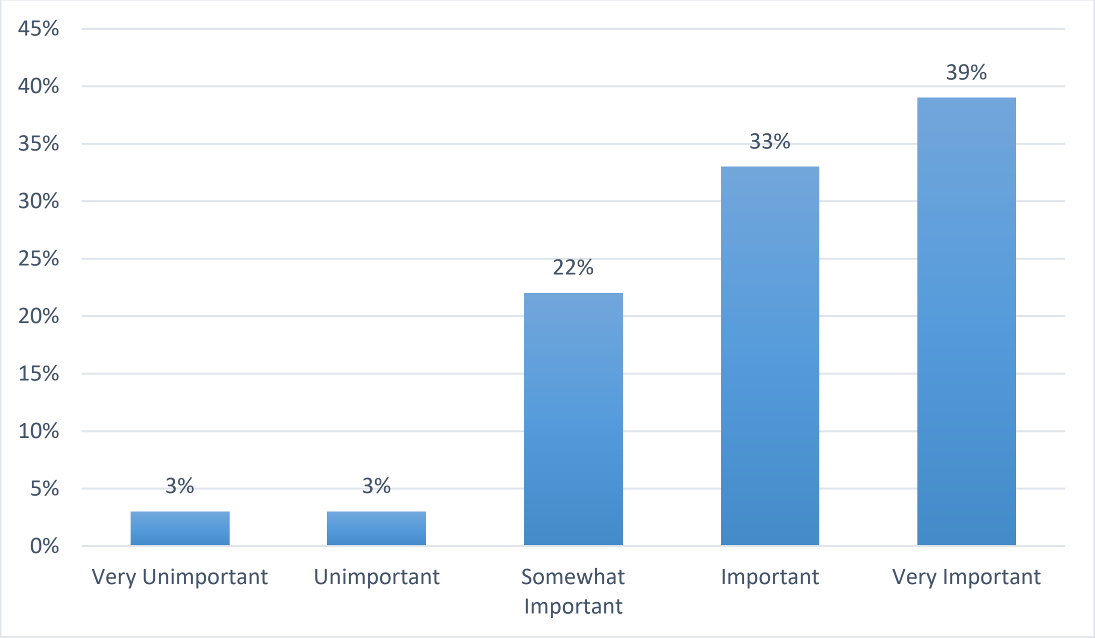


Figure 6.2: Respondents' views about the importance of Constitutive Acts

A high percentage of respondents (about 33%) chose 'very important' for having Functional Acts in Action-based Learning scenarios, and 31% of respondents chose 'important' (Figure 6.3).

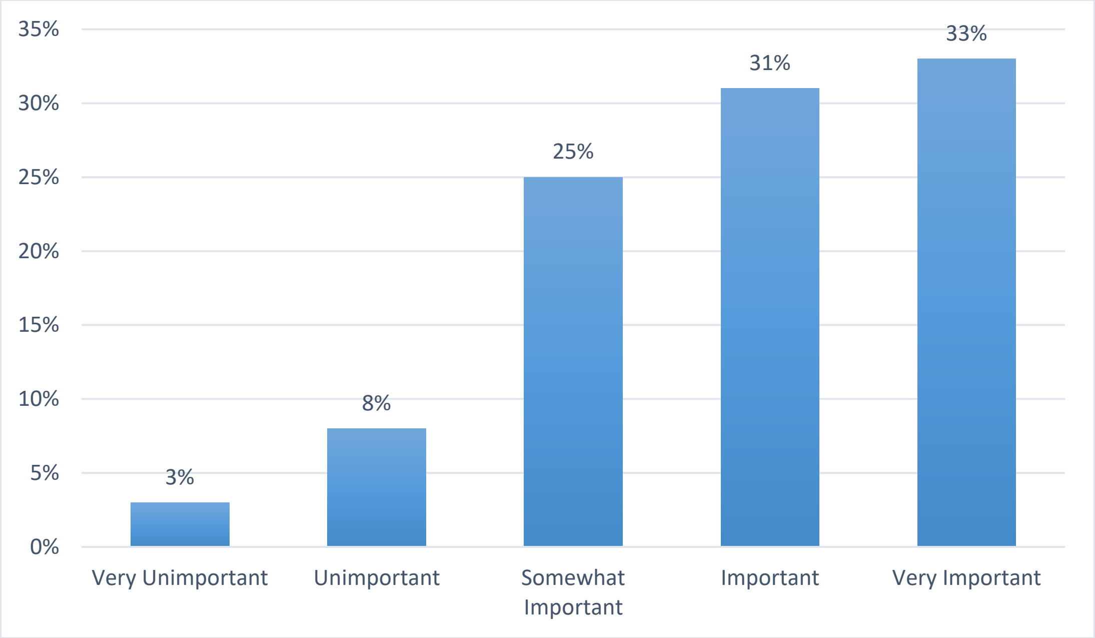


Figure 6.3: Respondents' views about the importance of Functional Acts

6.2.1.1.1. Friedman's test for prioritising the taxonomy's levels

The Friedman's test is used to investigate the priority that respondents give to the various levels of the taxonomy.

Null hypothesis (H0): There is no difference between the importance of the three levels.

Alternative hypothesis (H1): There is a difference between the importance of the three levels.

Table 6.7: Friedman's test's descriptive statistics

Ranks	
	Mean Rank
Q1	2.43
Q2	1.86
Q3	1.71

Table 6.8: Friedman's test's results

Test Statisticsa	
N	36
Chi-Square	18.317
Df	2
Asymp. Sig.	.000
a. Friedman Test	

According to the resulting Asymptotic Significance of zero (Sig. = 0.000), we can conclude that there is a clear difference between the three levels so the null hypothesis is rejected and the alternative is accepted; there are differences among the importance of the three levels. The highest rate of 2.43 (Table 6.7) goes to the Goal Act; the second highest rank belongs to the Constitutive Acts with 1.86 and lastly, there is the Functional Acts ranked as 1.71.

6.2.1.2. The Importance of Functional Classes, BEHAVE, Multiple Expert Solutions, And Formative Feedback

The following hypotheses are considered for the questions 4 to 14:

Null hypothesis (H0): Respondents have no tendency towards the importance of the subject of Qi.

Alternative hypothesis (H1): Respondents have tendency towards the importance of the subject of Qi.

i = 4,…, 14;

Table 6.9 (See Appendix 14) shows the descriptive statistics of the survey. The respondents' responses to questions 4 to 14 can be seen in Figures 6.4 – 6.14 (See Appendix 14). Table 6.10 demonstrates the Chi-square statistics and results of the survey, indicating the acceptance or the rejection of the hypothesis for each question.

Table 6.10: Chi-square statistics and results for questions 4 to 14

		Chi-square	df	Asymp. Sig.	Result
Q4	How important is it to recognise the locomotion of the avatars as an action, to assess the trainees' performance in any 'Action-based Learning' scenario?	16.222	4	0.003	RH0
Q5	How important is it to recognise different types of locomotive actions in any 'Action-based Learning' scenario?	20.389	4	0.000	RH0
Q6	How important is it to recognise the trainees' response as an actions to assess the trainees' performance in any 'Action-based Learning' scenario? *	20.389	4	0.000	RH0
Q7	How important is it to recognise the difference of trainees' response from actions with different purposes? * (Pushing the button with the green light vs. pushing a button)	20.389	4	0.000	RH0
Q8	How important is it to recognise those trainees' actions which are reflecting their decisions, to assess the trainees' performance in any 'Action-based Learning' scenario?	15.333	4	0.002	RH0
Q9	How important is it to recognise the operative actions of trainees to assess their performance in any 'Action-based Learning' scenario?	28.444	4	0.000	RH0
Q10	How important is it to differentiate the actions which are not changing the objects (like breaking them) or the environment with the actions that are changing the structure or the nature of the objects?	33.722	4	0.000	RH0
Q11	How important is it to recognise the gestures of avatars as a communication method, to assess the trainees' performance in any 'Action-based Learning' scenario?	10.667	4	0.031	RH0
Q12	How important is it to assess the trainees' performance based on multiple experts' solutions instead of one expert?	15.389	4	0.004	RH0
Q13	Most assessments are performed by one assessor, but the Action-based Learning Assessment method uses a panel of assessors instead. What number of experts do you think is more suitable as a panel?	31.778	3	0.000	RH0
Q14	How important is to have a clear classification of actions in Action-based Learning Assessment?	27.056	4	0.000	RH0

Based on the information in the descriptive statistics table, there is a clear difference between the observed and expected values, which is a significant difference as the values of p-value (Asymp. Sig.) in questions 4 to 14 are all less than 0.05; hence, the null hypothesis is rejected. Therefore, the Chi-square test confirms the differences shown in the descriptive statistics table. As can be seen in the bar charts, the respondents evaluated the subjects in questions 4 to 14 as 'important'. Section 7.2.1 discusses these results in detail.

6.2.1.3. Known Current Taxonomies of Human Actions

Respondents are asked about their level of familiarity with taxonomies. They are also asked to provide the name of taxonomies of human actions used for Action-based Learning and assessment in 3D VTEs. The degree of familiarity of the respondents with taxonomies is illustrated in Figure 6.15.

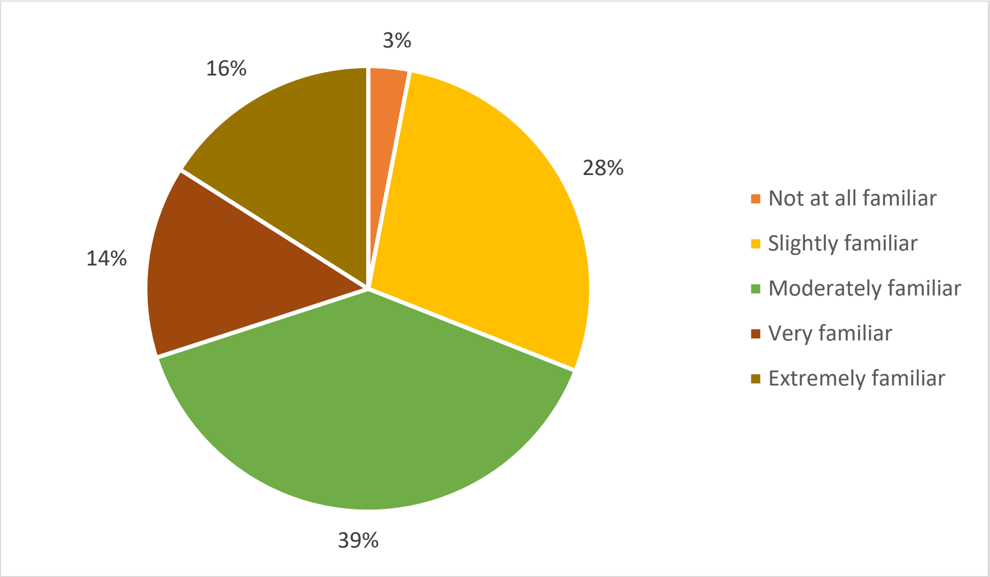


Figure 6.15: Distribution of respondents' views about their level of familiarity with the use of taxonomy.

The binomial test is used when a certain segment needs to be investigated. In investigating the equal number of respondents knowing any taxonomies of human actions for Action-based Learning and assessment in 3D VTE, the binomial test is used.

$$\begin{cases} H_0 : p = 0.5 \\ H_1 : p \neq 0.5 \end{cases}$$

As can be seen in Table 6.11, Sig is less than 0.05, so the null hypothesis is rejected. The majority of respondents did not know any taxonomy of human actions in 3D VTE.

Table 6.11: Binomial test results for all respondents

Binomial Test						
		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
Do you know any taxonomy of human actions for Action-based Learning and assessment in virtual training environments?*	Group 1	No	32	.89	.50	.000
	Group 2	Yes	4	.11		
	Total		36	1.00		

The four respondents who answered ‘yes’ to Question 16 (Do you know any taxonomy of human actions for Action-based Learning and assessment in virtual training environments?) indicating that they know a taxonomy of human actions for Action-based Learning and assessment in 3D VTE, gave the following responses in the open text field of the survey:

- one did not mention any taxonomy;
- two mentioned the (BEHAVE/Goal Oriented Actions) taxonomy of human actions[1];
- one mentioned the Karam & Schraefel (2005) taxonomy of gestures in Human Computer Interaction, which was considered in the literature review (Appendix 13).

6.2.2. The Results of VW and Industry Experts

To determine whether there is any difference between the views of virtual worlds’ experts and industry experts, the results of both groups are presented in this section.

6.2.2.1. Virtual World Experts: Questions 1-3

The Cronbach’s Alpha reliability coefficient for this group is 0.885 (See Appendix 14) which satisfies the criterion of Cronbach’s Alpha reliability coefficient having to be higher than 0.7 to be acceptable. The Chi-square test is used to determine whether the respondents have any tendency towards the importance of the three levels of the BEHAVE taxonomy.

6.2.2.1.1. Chi-square goodness of fit test for questions 1-3:

Null hypothesis (H0): Respondents have no tendency towards the importance of the subject of Qi.

Alternative hypothesis (H1): Respondents have a tendency towards the importance of the subject of Qi.

i = 1,…, 3;

[1] None of the participants was related to this research or the researcher. At the time the BEHAVE taxonomy had been presented at SLACTIONS 2013 and had been published in Fardinpour & Reiners (2014).

As mentioned in Section 6.2.1, the comparison of the expected and observed frequencies (Tables 6.13, 6.14, and 6.15) shows a clear difference in choices among the respondents.

Table 6.13: Chi-square test descriptive statistics for the Goal Act

How important is it to have a clear goal to achieve in any 'Action-based Learning' scenario?*			
	Observed N	Expected N	Residual
somewhat important	4	6.0	-2.0
important	2	6.0	-4.0
very important	12	6.0	6.0
Total	18		

Table 6.14: Chi-square test descriptive statistics for the Constitutive Acts

How important is it to break down the solution into different Constitutive Acts? *			
	Observed N	Expected N	Residual
very unimportant	1	3.6	-2.6
unimportant	1	3.6	-2.6
somewhat important	3	3.6	-.6
important	4	3.6	.4
very important	9	3.6	5.4
Total	18		

Table 6.15: Chi-square test descriptive statistics for the Functional Acts

How important is it to break down the milestones into Functional Acts? *			
	Observed N	Expected N	Residual
very unimportant	1	3.6	-2.6
unimportant	3	3.6	-.6
somewhat important	6	3.6	2.4
important	5	3.6	1.4
very important	3	3.6	-.6
Total	18		

Tables 6.16 and 6.17 show the Chi-square test outputs for Questions 1 and 2 including Chi-square statistic and Asymptotic Significance. For both questions, the Asymptotic Significance is less than 0.05 significance level, so the null hypothesis is rejected.

Table 6.16: Chi-square test statistics and results.

Test Statistics How important is it to have a clear goal to achieve in any 'Action-based Learning' scenario?*	
Chi-Square	9.333a
df	2
Asymp. Sig.	.009
a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 6.0.	

Table 6.17: Chi-square test statistics and results.

Test Statistics	
How important is it to break down the solution into different Constitutive Acts? *	
Chi-Square	12.000a
df	4
Asymp. Sig.	.017
a. 5 cells (100.0%) have expected frequencies less than 5. The minimum expected cell frequency is 3.6.	

For Question 1, chi-square test confirms the differences seen in the descriptive statistics table and shows that high percentage of respondents (about 67%) chose ‘very important’ for having the Goal Act in Action-based Learning scenarios, and 11% of respondents chose ‘important’ (Figure 6.16). Moreover, chi-square test for Question 2, confirms the differences seen in the descriptive statistics table and shows that a high percentage of respondents (about 50%) chose ‘very important’ for having Constitutive Acts in Action-based Learning scenarios, and 22% of respondents chose ‘important’ (Figure 6.17).

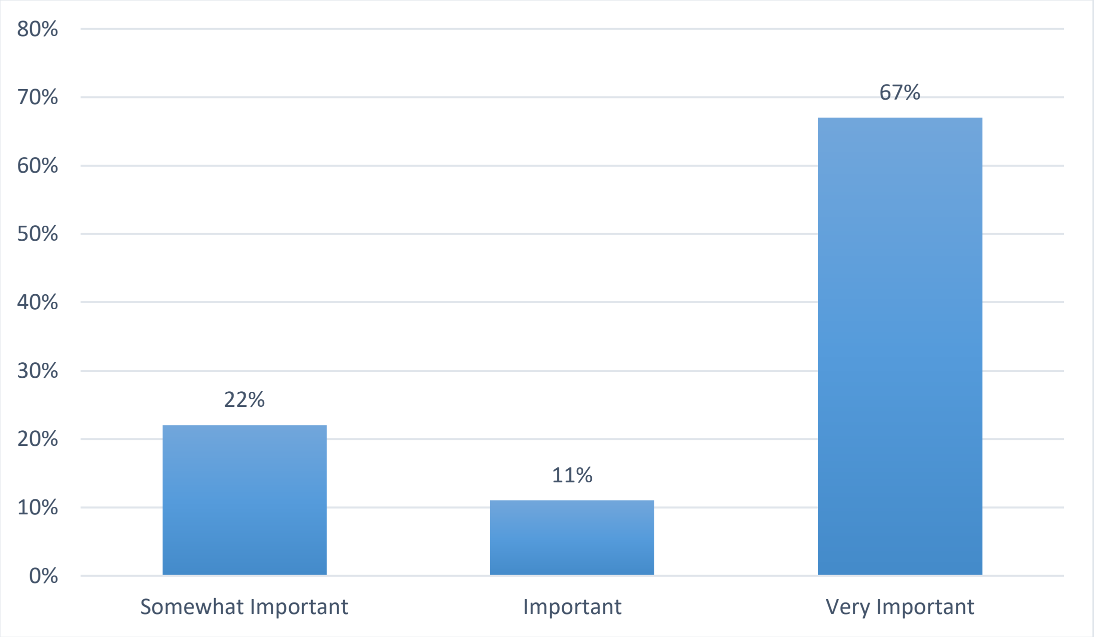


Figure 6.16: Respondents’ views about the importance of the Goal Act

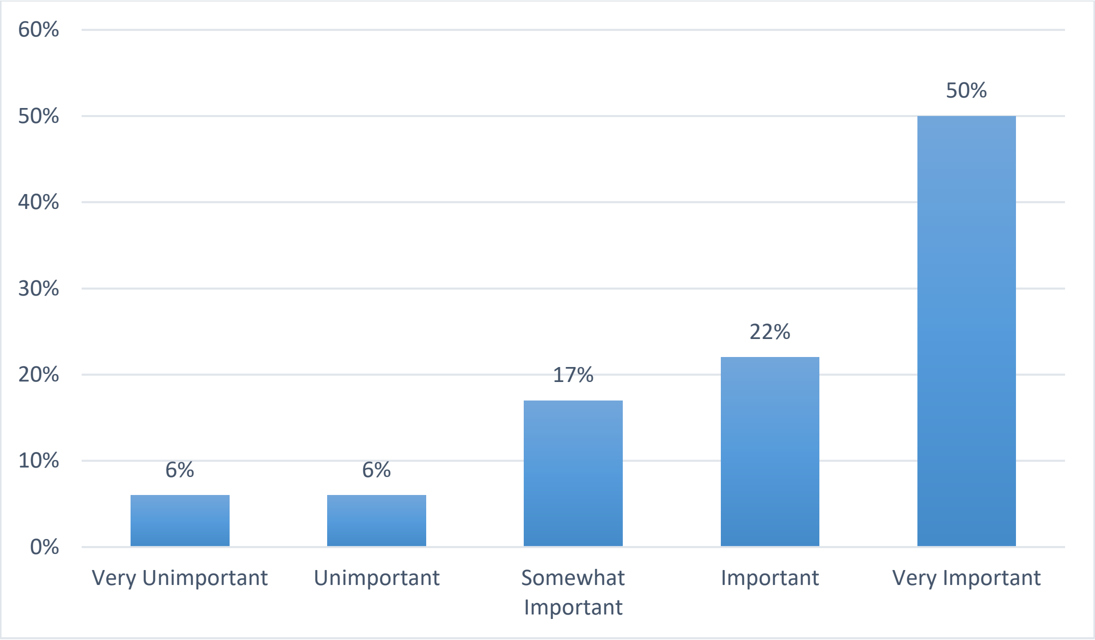


Figure 6.17. Respondents' views about the importance of Constitutive Acts

However, as shown in Table 6.18, the p-value (Asymp. Sig.) is equal to 0.377 which is greater than 0.05 significance level, so the null hypothesis that indicates the absence of tendency towards the importance of Functional Acts in Action-based Learning scenarios is accepted.

Table 6.18: Chi-square test statistics and results.

Test Statistics	
	How important is it to break down the milestones (in any 'Action-based Learning' scenario) into Functional Acts? *
Chi-Square	4.222a
df	4
Asymp. Sig.	.377
a. 5 cells (100.0%) have expected frequencies less than 5. The minimum expected cell frequency is 3.6.	

Therefore, the Chi-square test shows that virtual worlds' experts do not concur on the importance of Functional Acts in Action-based Learning scenarios (Figure 6.18).

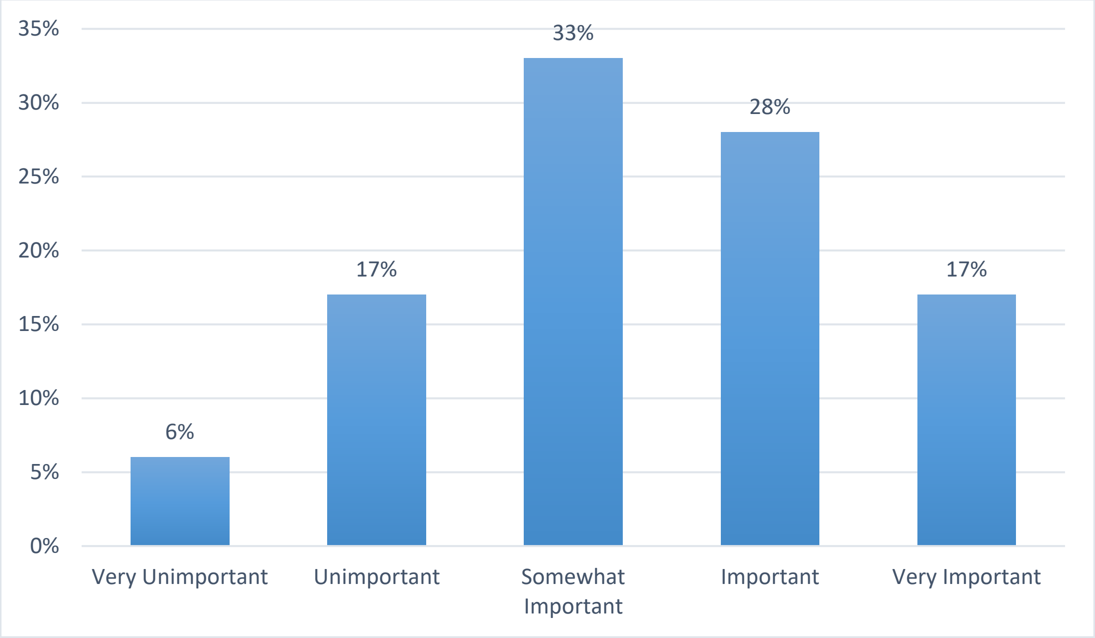


Figure 6.18. Respondents' views about the importance of Functional Acts

6.2.2.1.2. Friedman's test for prioritising the taxonomy levels

To investigate the priority of the taxonomy's levels among the respondents, the Friedman's test is used.

Null hypothesis (H0): There is no difference in importance between the three levels.

Alternative hypothesis (H1): There is a difference in importance between the three levels.

According to the resulting Asymptotic Significance of 0.001 (Table 6.20), we can conclude that there is a clear difference between the three levels, so the null hypothesis is rejected. The highest rank of 2.44 (Table 6.19) goes to the Goal Act; the second highest rank belongs to the Constitutive Acts with 2.11, and the Functional Acts have a rank of 1.44.

Table 6.19: Friedman's test's descriptive statistics

Ranks	
	Mean Rank
Q1	2.44
Q2	2.11
Q3	1.44

Table 6.20: Friedman test's results

Test Statisticsa	
N	18
Chi-Square	13.440
Df	2
Asymp. Sig.	.001
a. Friedman Test	

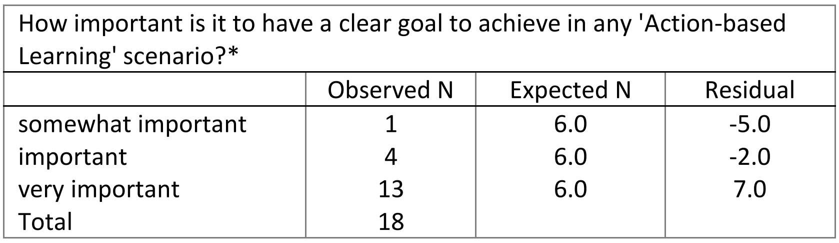
6.2.2.2. Industry Experts: Questions 1-3

The same tests are used for the real-life industry experts. The Cronbach's Alpha reliability coefficient is 0.891 which satisfies the criterion of Cronbach's Alpha reliability coefficient having to be higher than 0.7 to be acceptable (See Appendix 14).

6.2.2.2.1. Chi-square goodness of fit test for questions 1-3:

The same hypotheses are used for this group as well. The results are illustrated in the following three tables.

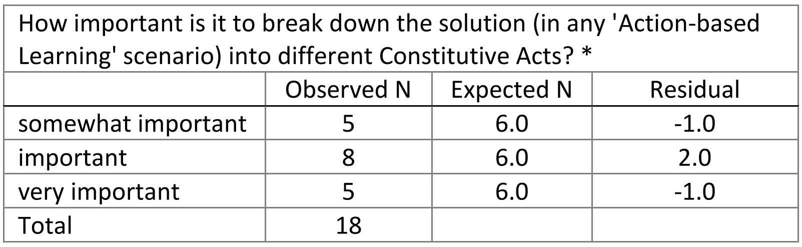
Table 6.21: Chi-square test descriptive statistics for the Goal Act



How important is it to have a clear goal to achieve in any 'Action-based Learning' scenario?*

	Observed N	Expected N	Residual
somewhat important	1	6.0	-5.0
important	4	6.0	-2.0
very important	13	6.0	7.0
Total	18		

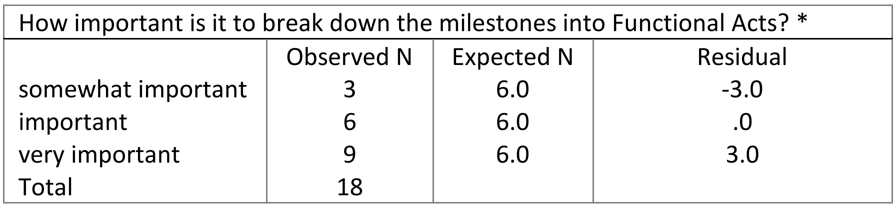
Table 6.22: Chi-square test descriptive statistics for the Constitutive Acts



How important is it to break down the solution (in any 'Action-based Learning' scenario) into different Constitutive Acts? *

	Observed N	Expected N	Residual
somewhat important	5	6.0	-1.0
important	8	6.0	2.0
very important	5	6.0	-1.0
Total	18		

Table 6.24: Chi-square test descriptive statistics for the Functional Acts



How important is it to break down the milestones into Functional Acts? *

	Observed N	Expected N	Residual
somewhat important	3	6.0	-3.0
important	6	6.0	.0
very important	9	6.0	3.0
Total	18		

In this group, respondents only showed the tendency towards the importance of the Goal Act in Action-based Learning scenarios (Question 1). Respondents had no tendency towards the importance of the Constitutive, and Functional Acts (Tables 6.25, 6.26, and 6.27).

Table 6.25: Chi-square test statistics and results

Test Statistics	
	How important is it to have a clear goal to achieve in any 'Action-based Learning' scenario?*
Chi-Square	13.000a
df	2
Asymp. Sig.	.002
a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 6.0.	

Table 6.26: Chi-square test statistics and results.

Test Statistics	
	How important is it to break down the solution (in any 'Action-based Learning' scenario) into different Constitutive Acts? *
Chi-Square	1.000a
df	2
Asymp. Sig.	.607
a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 6.0.	

Table 6.27: Chi-square test statistics and results.

Test Statistics	
	How important is it to break down the milestones (in any 'Action-based Learning' scenario) into Functional Acts? *
Chi-Square	3.000a
df	2
Asymp. Sig.	.223
a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 6.	

Chi-square test confirms the differences seen in the descriptive statistics in the Figure 6.19 that shows that a high percentage of respondents (about 72%) chose ‘very important’ for having the Goal Act in Action-based Learning scenarios, and 22% of respondents chose ‘important’.

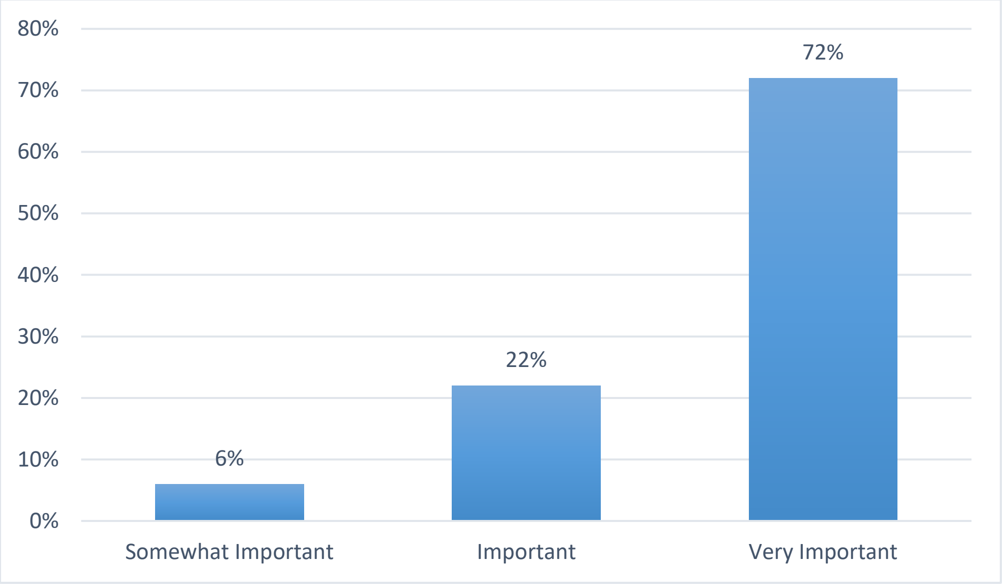


Figure 6.19: Respondents' views about the importance of the Goal Act

However, the Chi-square test shows that industry experts do not have a unanimous opinion about the importance of Constitutive Acts in Action-based Learning scenarios although, as can be seen in Figure 6.20, 44% of them believed it is 'important', and 28% equally believed it is 'somewhat important' and 'very important'.

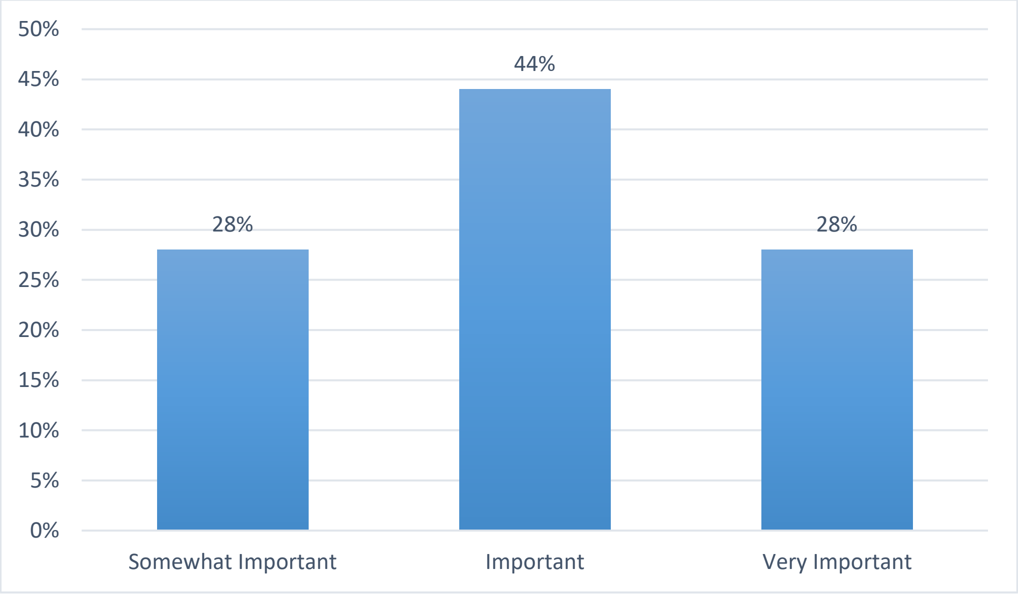


Figure 6.20. Respondents' views about the importance of Constitutive Acts

Similarly, the Chi-square test shows that industry experts do not have a unanimous opinion about the importance of Functional Acts in Action-based Learning scenarios, although 50% of them believed it is 'very important' and 33% 'important' (Figure 6.21).

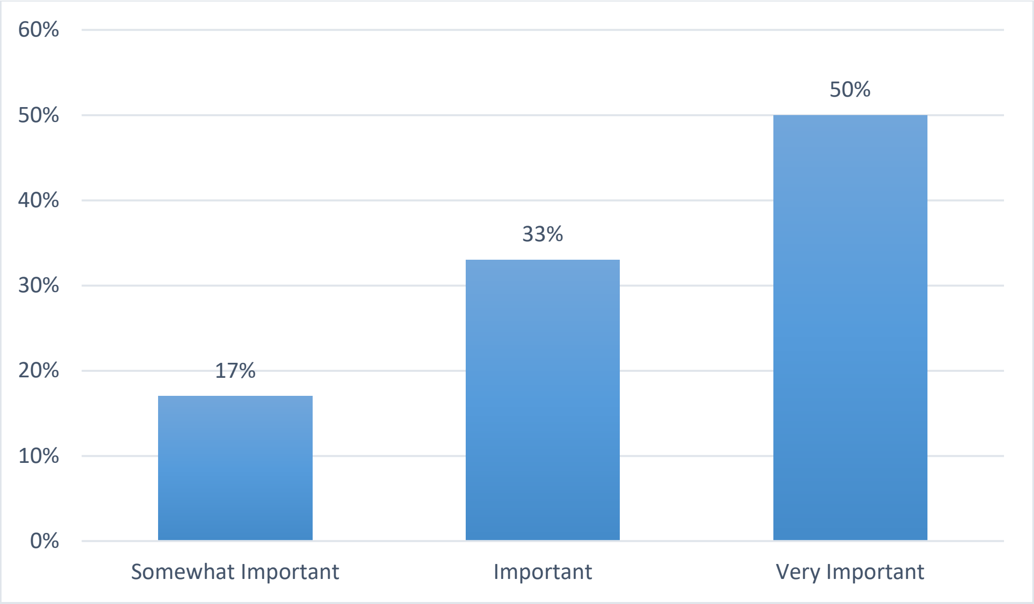


Figure 6.21. Respondents’ views about the importance of Functional Acts

6.2.2.2.2. Friedman's test for prioritising the taxonomy levels

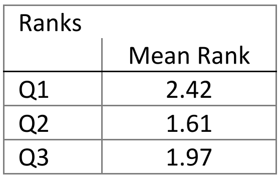
To investigate the priority of the taxonomy’s levels among the respondents, the Friedman’s test is used.

Null hypothesis (H0): There is no difference in importance between the three levels.

Alternative hypothesis (H1): There is a difference in importance between the three levels.

According to the resulting Asymptotic Significance of 0.001, we can conclude that there is a clear difference between the three levels, so the null hypothesis is rejected. The highest rate of 2.42 (Table 6.28) goes to the Goal Act; the second highest rank belongs to the Constitutive Acts with 1.97, and the Functional Acts have the rank 1.61. Chi-square statistic is equal to 13.188 (Table 6.29) with a degree of freedom of two.

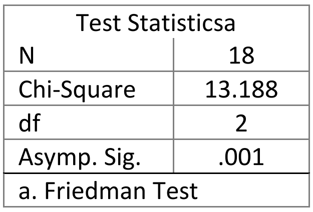
Table 6.28: Friedman’s test’s descriptive statistics



Ranks

	Mean Rank
Q1	2.42
Q2	1.61
Q3	1.97

Table 6.29: Friedman test’s results



Test Statisticsa

N	18
Chi-Square	13.188
df	2
Asymp. Sig.	.001

a. Friedman Test

6.2.2.3. Both Groups' Results for the Questions 4-16

While the gathered and analysed views of respondents regarding the importance of the subjects of questions 4 to 14, show a high degree of importance, it is important to compare the results from both groups that is those of the industry experts and the VW experts, to determine whether there are any differences between the two. An investigation of these differences yielded the following results.

Null hypothesis (H0): Respondents of Gi have no tendency towards the importance of the subject of Qj.

Alternative hypothesis (H1): Respondents of Gi have a tendency towards the importance of the subject of Qj.

i= 1, 2 and j = 4,… , 14;

Table 6.30 (See Appendix 14) shows the descriptive statistics for the industry experts. The descriptive statistics for the VW experts can be seen in Table 6.31 (See Appendix 14). Figures 5.22 – 5.32 (See Appendix 14) illustrate the comparison of both groups of respondents' responses to questions 4 to 14.

Table 6.32: Chi-square statistics and results of industry experts.

		Chi-square	df	Asymp. Sig.	Result
Q4	How important is it to recognise the locomotion of the avatars as an action, to assess the trainees' performance in any 'Action-based Learning' scenario?	8.222	3	0.042	RH0
Q5	How important is it to recognise different types of locomotive actions in any 'Action-based Learning' scenario?	7.778	3	0.051	RH0
Q6	How important is it to recognise the trainees' response as an actions to assess the trainees' performance in any 'Action-based Learning' scenario? *	5.556	3	0.135	AH0
Q7	How important is it to recognise the difference of trainees' response from actions with different purposes? * (Pushing the button with the green light vs. pushing a button)	9.111	3	0.028	RH0
Q8	How important is it to recognise those trainees' actions which are reflecting their decisions, to assess the trainees' performance in any 'Action-based Learning' scenario?	0.889	1	0.346	AH0
Q9	How important is it to recognise the operative actions of trainees to assess their performance in any 'Action-based Learning' scenario?	7.0	2	0.030	RH0
Q10	How important is it to differentiate the actions which are not changing the objects (like breaking them) or the environment with the actions that are changing the structure or the nature of the objects.	16.333	2	0.000	RH0
Q11	How important is it to recognise the gestures of avatars as a communication method, to assess the trainees' performance in any 'Action-based Learning' scenario?	10.0	3	0.019	RH0
Q12	How important is it to assess the trainees' performance based on multiple experts' solutions instead of one expert?	7.333	3	0.062	AH0
Q13	Most assessments are performed by one assessor, but the Action-based Learning Assessment method uses a panel of assessors instead. What number of experts do you think is more suitable as a panel?	18.0	3	0.000	RH0
Q14	How important is to have a clear classification of actions in Action-based Learning Assessment?	7.0	2	0.030	RH0

Chi-square test statistics and results of responses from industry experts can be found in Table 6.32. In Table 6.32, the p-value for questions 7, 9, and 13 is greater than 0.05, so the null hypothesis is accepted. For the rest of the questions that p-value is less than 0.05, the null hypothesis is rejected, which means in questions 6, 8, and 12 experts have no tendency towards the importance of the subject of these questions. However, in the rest of the questions, there is a preference of choice. The majority of industrial experts think that the subject of questions 5, 7, 9, 10, 11, and 14 are important and 12 out of 18 think that the best number of panel assessors is 1-3, which is a meaningful tendency.

Table 6.33: Chi-square statistics and results of VW experts.

		Chi-Square	df	Asymp. Sig.	Result
Q4	How important is it to recognise the locomotion of the avatars as an action, to assess the trainees' performance in any 'Action-based Learning' scenario?	13.111	4	0.011	RH0
Q5	How important is it to recognise different types of locomotive actions in any 'Action-based Learning' scenario?	9.778	4	0.044	RH0
Q6	How important is it to recognise the trainees' response as an actions to assess the trainees' performance in any 'Action-based Learning' scenario? *	4.667	3	0.198	AH0
Q7	How important is it to recognise the difference of trainees' response from actions with different purposes? * (Pushing the button with the green light vs. pushing a button)	2.889	3	0.409	AH0
Q8	How important is it to recognise those trainees' actions which are reflecting their decisions, to assess the trainees' performance in any 'Action-based Learning' scenario?	3.778	3	0.286	AH0
Q9	How important is it to recognise the operative actions of trainees to assess their performance in any 'Action-based Learning' scenario?	9.222	4	0.056	AH0
Q10	How important is it to differentiate the actions which are not changing the objects (like breaking them) or the environment with the actions that are changing the structure or the nature of the objects?	8.667	4	0.070	AH0
Q11	How important is it to recognise the gestures of avatars as a communication method, to assess the trainees' performance in any 'Action-based Learning' scenario?	3.111	4	0.539	AH0
Q12	How important is it to assess the trainees' performance based on multiple experts' solutions instead of one expert?	8.111	4	0.088	AH0
Q13	Most assessments are performed by one assessor, but the Action-based Learning Assessment method uses a panel of assessors instead. What number of experts do you think is more suitable as a panel?	7.0	2	0.030	RH0
Q14	How important is to have a clear classification of actions in Action-based Learning Assessment?	8.111	4	0.088	AH0

Table 6.33 shows the Chi-square test statistics and results of responses from VW's experts. In this table, the p-value for questions 6, 7, 8, 9, 10, 11, 12, and 14 is

greater than 0.05, so the null hypothesis is accepted. The p-value for questions 4, 5, and 13 is less than 0.05; that means the null hypothesis is rejected. This means, that for questions 6, 7, 8, 9, 10, 11, 12 and 14, experts have no tendency towards the importance of the subject of these questions but in questions 4, 5, and 13, there is a preference of choice. The majority of VW experts believe that the subjects of questions 4, 5, and 13 are ‘somewhat important’, ‘important’, or ‘very important’, and 10 out of 18 believe that the best number of panel assessors is 1-3, which is a meaningful tendency.

The bar chart in Figure 6.33 clearly shows the high degree of familiarity with taxonomies among the virtual worlds’ experts. In investigating the equal number of respondents knowing any taxonomies of human actions in 3D VTE, the binomial test is used.

$$\begin{cases} H_0 : p = 0.5 \\ H_1 : p \neq 0.5 \end{cases}$$

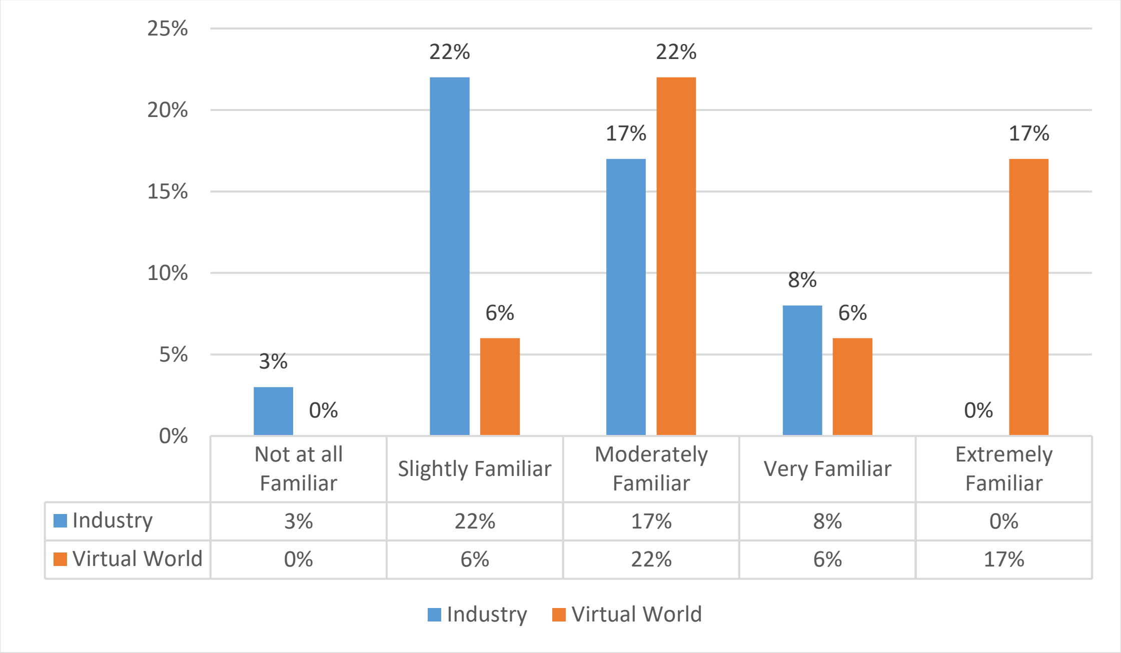


Figure 6.33: Comparing the distribution of respondents' level of familiarity with the use of human actions taxonomy.

As can be seen in Table 6.34, the p-value (Sig) is less than 0.05 for both groups so that the null hypothesis is rejected, and the majority of respondents did not know any taxonomy of human actions for Action-based Learning and assessment in 3D VTE. None of the experts in the industry experts group knew any taxonomy of human actions for Action-based Learning and assessment in 3D VTE, the same as the 78% of the VW experts.

Table 6.34: Binomial test results comparison between both groups

Binomial Test							
		Group	Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
RL	Do you know any taxonomy of human actions for Action-based Learning and assessment in virtual training environments?*	Group 1	No	18	1.00	.50	.000
		Total		18	1.00		
VE	Do you know any taxonomy of human actions for Action-based Learning and assessment in virtual training environments?*	Group 1	No	14	.78	.50	.031
		Group 2	Yes	4	.22		
		Total		18	1.00		

6.2.2.3.1. Chi-Square Test of Homogeneity

The homogeneity test is used to compare the views of two groups of respondents. This is used to discover whether the respondents in the two groups have similar opinions about the survey’s questions.

Null hypothesis (H0): both groups of respondents answered the questions similarly.

Alternative hypothesis (H1): both groups of respondents answered the questions differently.

For the Chi-square test of homogeneity to be used, in the descriptive tables resulting from the test, cells with a frequency less than five should be merged so that none of the cells has a frequency less than five.

The results of the homogeneity test for Q1 – Q16 are shown in Table 6.35. Any result shown as AH0 means that the Null hypothesis is accepted, and RH0 means that the Null hypothesis is rejected. However, because the contingency table for the Q16 is 2*2 (the question is a Yes/No question), and the observed frequency of two cells is less than five and cannot be merged, the p-value resulting from the Exact Fisher’s Test is used.

Table 6.35: Chi-Square Test of Homogeneity results

		Chi-square	df	Asymp. Sig.	Result
Q1	How important is it to have a clear goal to achieve in any 'Action-based Learning' scenario?	0.131	1	0.717	AH0
Q2	How important is it to break down the solution (in any 'Action-based Learning' scenario) into different milestones?	2.476	2	0.290	AH0
Q3	How important is it to break down the milestones (in any 'Action-based Learning' scenario) into basic actions?	6.860	2	0.032	RH0
Q4	How important is it to recognise the locomotion of the avatars as an action, to assess the trainees' performance in any 'Action-based Learning' scenario?	8.042	2	0.018	RH0
Q5	How important is it to recognise different types of locomotive actions in any 'Action-based Learning' scenario?	2.143	2	0.343	AH0
Q6	How important is it to recognise the trainees' response as an actions to assess the trainees' performance in any 'Action-based Learning' scenario? *	0.619	2	0.734	AH0
Q7	How important is it to recognise the difference of trainees' response from actions with different purposes? * (Pushing the button with the green light vs. pushing a button)	2.220	2	0.330	AH0
Q8	How important is it to recognise those trainees' actions which are reflecting their decisions, to assess the trainees' performance in any 'Action-based Learning' scenario?	9.327	2	0.009	RH0
Q9	How important is it to recognise the operative actions of trainees to assess their performance in any 'Action-based Learning' scenario?	4.400	2	0.111	AH0
Q10	How important is it to differentiate the actions which are not changing the objects (like breaking them) or the environment with the actions that are changing the structure or the nature of the objects?	9.600	2	0.008	RH0
Q11	How important is it to recognise the gestures of avatars as a communication method, to assess the trainees' performance in any 'Action-based Learning' scenario?	6.452	2	0.040	RH0
Q12	How important is it to assess the trainees' performance based on multiple experts' solutions instead of one expert?	3.347	2	0.188	AH0
Q13	Most assessments are performed by one assessor, but the Action-based Learning Assessment method uses a panel of assessors instead. What number of experts do you think is more suitable as a panel?	0.000	1	1.000	AH0
Q14	How important is to have a clear classification of actions in Action-based Learning Assessment?	6.322	2	0.042	RH0
Q15	To what extent are you familiar with the use of taxonomies?	7.013	2	0.030	RH0
Q16	Do you know any taxonomy of human actions for Action-based Learning and assessment in virtual training environments?	4.500	1	0.104	AH0

Because the p-value for questions 3, 4, 8, 10, 11, 14 and 15, is less than 0.05, the H0 is rejected; hence, the views are not similar, and therefore there is a significant difference in the points of view expressed by the two groups.

6.3. Card Sorting Test Results

As explained in Section 3.6.2, the card sorting test is used to evaluate the exhaustiveness of the Functional classes. In this validity test, 255 people participated, and 207 (81%) of those people sorted all 47 cards. The card sorting test system reported an average time of five minutes for the test, indicating that the majority of the respondents were able to use the classes easily. The following tests are used to analyse the data from the card sorting test.

6.3.1. Cluster Analysis

K-mean cluster analysis (Everitt et al., 2011) is used to analyse the sorted data taking into account that the clustering method is partitional (non-overlapping clusters), items should be assigned to clusters with closer centres, and the number of clusters is specified. SPSS software is used to standardise the weighted proximity matrix[1] provided by the Syncaps cluster analysis software (Spencer, 2009). Before analysing the data, it is helpful to create the graphical views of the multivariate data as they help to understand the structure of the data (Everitt et al., 2011).

Scatterplots of each pair of variables, arranged as a scatterplot matrix, can be used to find cluster structures in data when we have multivariate data.

> “A scatterplot matrix is defined as a square, symmetric grid of bivariate scatterplots. This grid has p rows and p columns, each one corresponding to a different one of the p variables. Each of the grid’s cells show a scatterplot of two variables. [] The scatterplot matrix is symmetric about its diagonal” (Everitt et al., 2011: p. 24).

The Scatterplot matrix can indicate: whether there are pairwise relationships between the variables; if so, what the nature of these relationships is; if there are outliers in the data; and if the data is clustered in groups (NCroarkin & Tobias, 2015). In the context of cluster analysis, if the scatterplot shows a clear linear or non-linear form and shape, and a relationship between the variables, the clusters might show an

[1] A proximity matrix is an n by n matrix comprising of all the pairwise dissimilarities or similarities between the sorted items.

association that undermines the exhaustive clusters condition. These shapes and forms are detectable when the data, represented with dots or circles, form a line or curve. At the same time, the relationships can be detected from the data represented by the density of the dots or circles concentrated in one direction.

This research uses the data (Shown in Table 6.45, Appendix 14) collected in the study for 47 cards. The recorded variables are Constructional, Decisional, Responsive, Gestural, Locomotive and Operative. The six proposed variables are used to assess whether there is any evidence that there are any outliers or clustering by groups in the data.

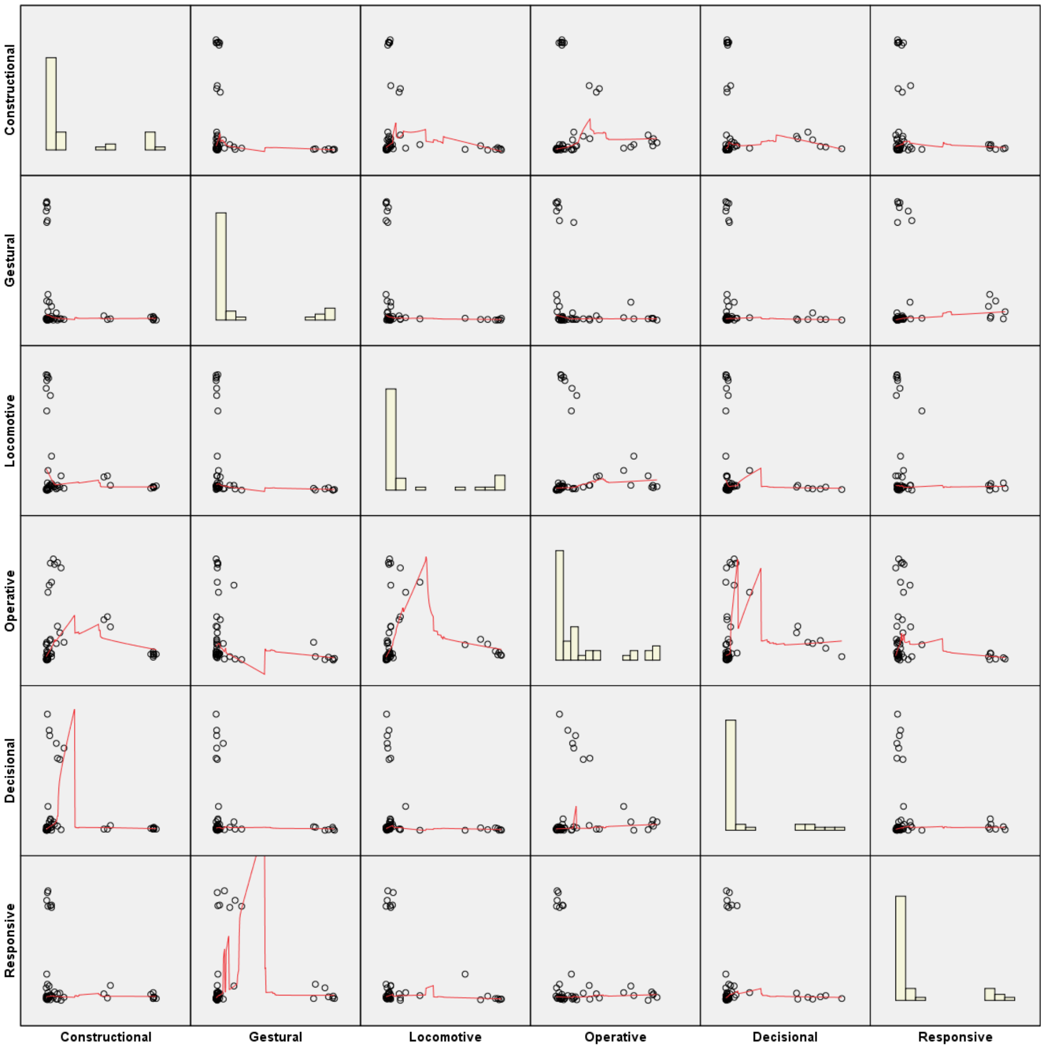


Figure 6.34: Scatterplots' matrix for card sorting test

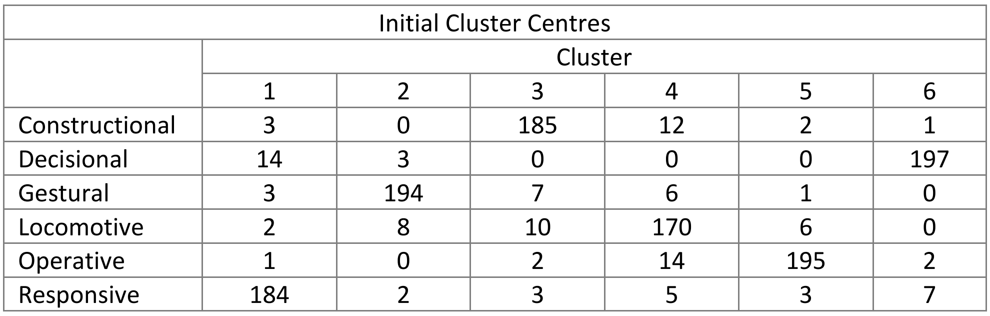
The scatterplot matrix in Figure 6.34 represents the various combinations of each variable plotted against others on the X and Y axes. The red Loess curve illustrates the nonparametric regression with a fit point (percent of population) of 75%

and Gaussian weight function (highest weight is assigned to the closes point). The diagonal illustrates the histogram of each plot.

An outlier is a data point that differs significantly from other observations in a set of data. The presence of outliers can change the result of the cluster analysis if the outliers form a new cluster within a cluster, comparable in size. In a scatterplot matrix, outliers can be recognised by their direction and distance from the rest of the data points. The scatterplot matrix is a useful visual tool for detecting any particular pattern or outliers in a multivariate data analysis. The results illustrated in Figure 6.33 are investigated in Section 7.3, to determine the possible associations (linear or curved relationship) between variables, and the possible outliers in any of the plots of the matrix.

The K-mean cluster analysis method is performed in four steps (Section 3.6.2), beginning with breaking up the data into K clusters (where K=6 in this case) and assigning K random points (one point to each cluster). Table 6.36 shows the initial cluster centres[1].

Table 6.36: Initial cluster centres for card sorting test



Initial Cluster Centres						
	Cluster					
	1	2	3	4	5	6
Constructional	3	0	185	12	2	1
Decisional	14	3	0	0	0	197
Gestural	3	194	7	6	1	0
Locomotive	2	8	10	170	6	0
Operative	1	0	2	14	195	2
Responsive	184	2	3	5	3	7

The second step consists of repeated runs to calculate the distance[2] (the distance is in n-dimensions space) of each item from its centre and then calculate the standard error of the mean. Table 6.37 (See Appendix 14) shows the changes to the cluster centres with each iteration. The third step is to optimise the K-mean solution in which the item furthest from the cluster centre is moved to a cluster closest to the centre. Finally, in the fourth step, the algorithm stops when the cluster members are no longer changing, that is, when the calculated value of the standard error of the mean is no

[1] The cluster centre is the value of the variables for an item. In Table 6.37, six items with the maximum value are chosen to form the initial cluster centres.

[2] Euclidean distance is the distance between the two points (in here: item and the cluster centre) in Euclidean space. The Euclidean distance in can be calculated by: $D_{ij}^{2} = \sum_{v=1}^{n}(X_{vi} - X_{vj})^{2}$

longer decreasing. Table 6.38 shows the final cluster centres that result after the stop command.

Table 6.38: Final cluster centres that result after the stop command

Final Cluster Centres						
	Cluster					
	1	2	3	4	5	6
Constructional	5	2	157	12	15	1
Decisional	18	1	3	6	2	185
Gestural	6	177	9	18	4	1
Locomotive	5	18	27	148	31	7
Operative	6	2	3	13	149	3
Responsive	167	8	8	11	5	10

Table 6.39 shows the distance between cluster centres. The minimum distance of two cluster centres is 180.245 for clusters four and five. Also, the maximum distance of two cluster centres is 254.144 for clusters two and six.

Table 6.39: Distances between Final Cluster Centres

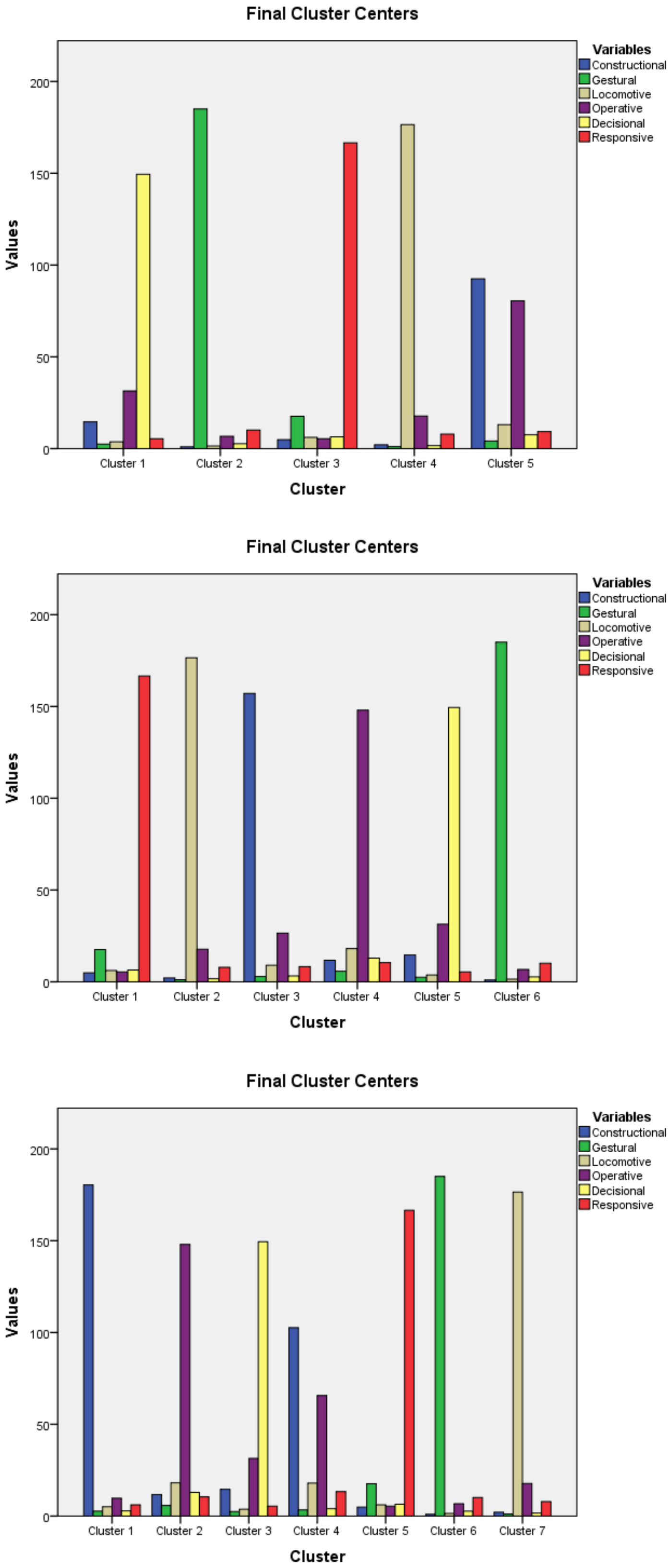
Distances between Final Cluster Centres						
Cluster	1	2	3	4	5	6
1	-	233.793	221.146	212.267	217.765	229.248
2		-	228.377	205.658	228.146	254.144
3			-	189.944	204.347	240.790
4				-	180.245	229.325
5					-	235.966
6						-

Table 6.40 shows which item belongs to what cluster and Table 6.41 (See Appendix 14) shows each item's distance from its cluster centre.

Table 6.40: Number of cases in each cluster

Number of Cases in each Cluster			Maximum distance from the cluster centre	Expert's Clusters
Cluster	1	7.000	28.29	Responsive
	2	8.000	57.21	Locomotive
	3	10.000	74.36	Constructional
	4	8.000	46.89	Operative
	5	7.000	53.86	Decisional
	6	7.000	32.37	Gestural
Valid		47.000		
Missing		.000		

The K-mean cluster analysis is also performed with five and seven clusters. As can be seen in Figure 6.35, by changing the number of clusters from six, to five or seven, at least, one of the cluster centres has changed, showing two close centres for one cluster. As a result, the six-cluster assumption in the K-mean cluster analysis is strongly accepted.



Final Cluster Centers

Final Cluster Centers

Final Cluster Centers

Figure 6.35: K-mean cluster analysis for five and seven clusters

6.3.2. Similarity of Calculated and Perceived Clusters

The internal validity of the clustering is examined by Fowlkes and Mallows index[1] in which the calculated clustering structure (C) (calculated in the K-mean cluster analysis test) and the perceived clustering structure (P) are compared, whereas P is based on expert's perception (Section 3.6.2), shown in Table 6.40 (Halkidi, Batistakis, and Vazirgiannis, 2002).

a is an observed pair of items which in the same cluster of structures C and P:

$$a = \binom{8}{2} + \binom{10}{2} + \binom{8}{2} + \binom{7}{2} + \binom{7}{2} + \binom{7}{2} = 28 + 45 + 28 + 21 + 21 + 21 = 164$$

b is an observed pair of items that are the same cluster of structure C but different clusters of structure P, which is equal to zero.

c is an observed pair of items that are in different clusters of structure C but in the same cluster of structure P, which is equal to zero.

d is an observed pair of items that are in different clusters of structure C and P, which is equal to zero.

$$\text{FM} = \sqrt{\frac{a}{a+b} \times \frac{a}{a+c}} = 1$$

In the Fowlkes and Mallows index, the value of one shows the maximum similarity between the calculated clustering structure (C) and the perceived clustering structure (P). The results are comparable with the 'popular placements matrix' whereby the most popular groups are indicated by the highest placement scores on each card. Each table cell shows the percentage of the respondents who sorted that card into the corresponding category (See Appendix 14, Table 6.45). Figure 6.36 (See Appendix 14) shows the colour map of the sorted cards.

6.3.3. Partitioning Validity: Dispersion of Data

R-squared[2] (R^2) index is used for internal validity of partitioning[3], the results of which can be seen in Table 6.42, Table 6.43, and Table 6.44 (See Appendix 14):

As a result:

[1] Fowlkes and Mallows index is used to measure the degree of similarity between the calculated clusters after a clustering algorithm, and the benchmark clusters by the experts.

[2] SSE is the sum of squares due to error and SST is the total sum of squares.

[3] Creating clusters that are not overlapping and each item only belongs to one specific cluster.

SST = $\sum_{i=1}^{n}\sum_{j=1}^{d}\left(X_{ij}-\overline{X}_{J}\right)^{2}$ = 997134.2

SSE = $\sum_{i=1}^{k}\sum_{x\in C_j}\sum_{j=1}^{d}\left(X_{ij}-\overline{X_{\iota J}}\right)^{2}$ = 46606.23

Moreover, R-squared index is equal to:

$R^2 = 1 - \frac{SSE}{SST} = \frac{SST- SSE}{SST} = \frac{997134.2-46606}{997134.2} = 0.95326$

The value range of this ratio is between zero and one. $R^2 = 0.95326$ is very close to one, which means that the dispersion of data inside the clusters is low; the distance between the clusters is high, which shows the high internal validity of the clustering.

6.3.4. Pearson Correlation Test

Also, the Pearson correlation test is used to investigate the correlation between the taxonomy classes. Pearson correlation can be positive, negative, or non-existent. Pearson values can vary between -1 and 1. Positive values represent positive linear correlation; negative values represent negative linear correlation; a value of 0 represents no linear correlation; and the closer the value is to 1 or –1, the stronger is the linear correlation. Evans & Over (2013) suggest that the absolute value of the Pearson correlation shows:

- 0.00-0.19: very weak correlation
- 0.20-0.39: weak correlation
- 0.40-0.59: moderate correlation
- 0.60-0.79: strong correlation
- 0.80-1.0: very strong correlation

To perform the Pearson correlation test two hypotheses are made:

Null hypothesis (H0): The classes are independent at the significance level of 0.05.

Alternative hypothesis (H1): The classes are dependent at the significance level of 0.05.

Table 6.47 shows the square matrix of the correlation coefficient between each variable with other five variables. Table 6.46 demonstrates correlation coefficients and precise p-value. For example, the correlation coefficient of Constructional with

Decisional equals – 0.268, and the 2-tailed p-value is 0.068, which is greater than 0.02, so the null hypothesis is accepted which indicates the independence of each class at the significance level of 0.05.

Table 6.46: Correlation matrix

Correlations							
		Constructional	Decisional	Gestural	Locomotive	Operative	Responsive
Constructional	Pearson Correlation	1	-.268	-.257	-.140	-.183	-.230
	Sig. (2-tailed)		.068	.082	.347	.218	.119
	N	47	47	47	47	47	47
Decisional	Pearson Correlation		1	-.261	-.284	-.210	-.083
	Sig. (2-tailed)			.077	.053	.156	.578
	N		47	47	47	47	47
Gestural	Pearson Correlation			1	-.129	-.237	-.211
	Sig. (2-tailed)				.386	.109	.154
	N			47	47	47	47
Locomotive	Pearson Correlation				1	-.032	-.262
	Sig. (2-tailed)					.830	.075
	N				47	47	47
Operative	Pearson Correlation					1	-.181
	Sig. (2-tailed)						.223
	N						47
Responsive	Pearson Correlation						1
	Sig. (2-tailed)						
	N						47

As can be seen in Table 6.46, the Pearson correlation test indicates that there is no relationship between any two pair of classes in the taxonomy, and the taxonomy has six exclusive classes, indicating the internal validity of the taxonomy. Furthermore, the results presented in Table 6.46 are applied to Evans & Over (2013)'s suggestion in Section 7.3.

6.4. Performance Coding Experiment Results

As explained in Section 3.6.3, the performance coding experiment is conducted to study the degree of similarity of the coded actions, using BEHAVE classification and syntax, with the real-life actions. The real-life performance is presented as a written scenario and a videotaped performance. The participants for the experiment are from Curtin University, and comprised students and staff in different departments, especially engineering and education (see Section 3.7 for the sampling method).

6.4.1. Normality Tests

Having 11 reference actions in written narration group and 14 reference actions in video group[1], the average of each group's participants' coded actions should be checked to see if there is a significant difference. The calculated mean for the number of actions for the video group is 9.30, and the variance is equal to 2.669 (Table 6.47).

Table 6.47: Descriptive statistics of number of actions in video group

One-Sample Statistics				
Video Results	N	Mean	Std. Deviation	Std. Error Mean
Number of actions	10	9.30	2.669	.844

To investigate the significance of the differences between the number of reference actions and determined actions by the participant, firstly we have to check for a normalised distribution. To study the normalisation hypothesis, the boxplot and normality test are used. Hence, the Kolmogorov-Smirnov normality test is used to find out if the samples are collected from a normal population.

The outlier value and median in the boxplot are other indicators of coincidence of the sampling distribution. There are no asymmetries or outlier values in the boxplot which means there is no reason for it not being a normal distribution (Figure 6.37).

[1] The reference performance was coded by the researcher given his expertise in machine tools and BEHAVE.

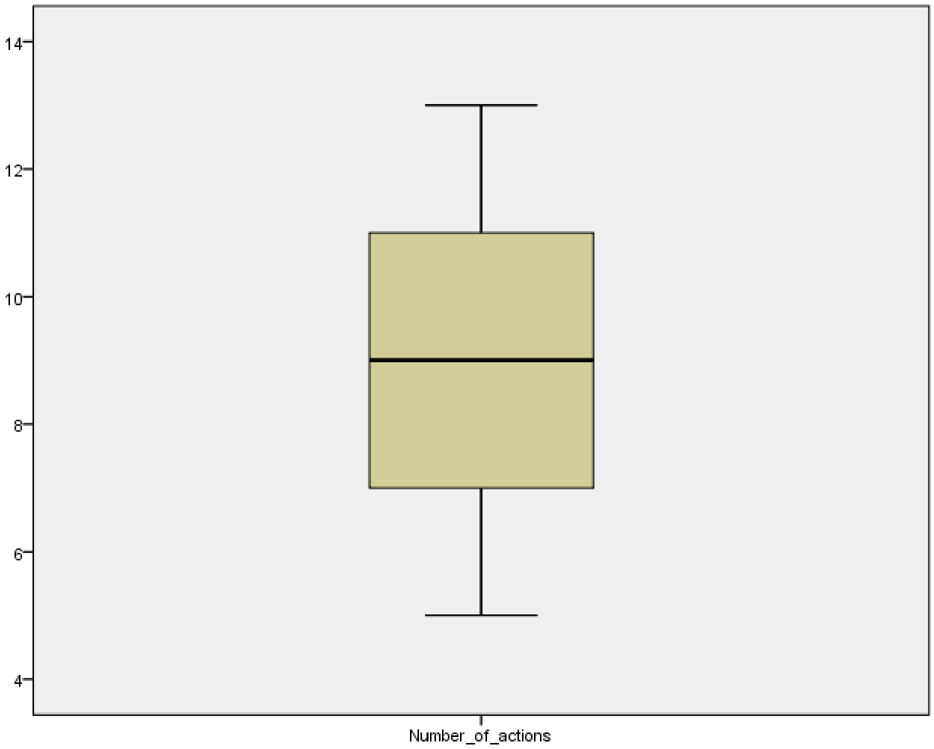
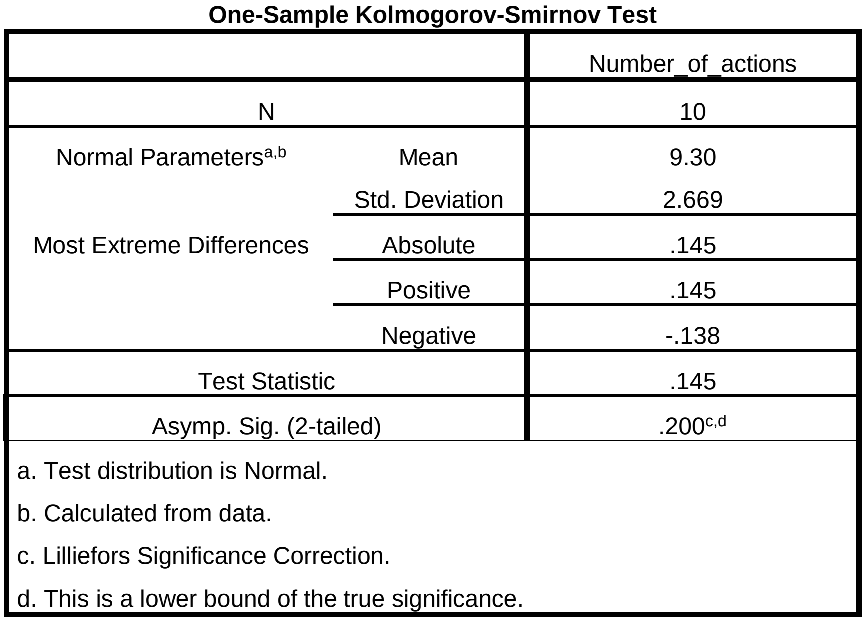


Figure 6.37: Boxplot of number of actions in video group

The Kolmogorov-Smirnov test results shown in Table 6.48 indicate that the population is normally distributed.

Table 6.48: Kolmogorov-Smirnov test results for video group



One-Sample Kolmogorov-Smirnov Test

		Number_of_actions
N		10
Normal Parameters[a,b]	Mean	9.30
	Std. Deviation	2.669
Most Extreme Differences	Absolute	.145
	Positive	.145
	Negative	-.138
Test Statistic		.145
Asymp. Sig. (2-tailed)		.200[c,d]

a. Test distribution is Normal.

b. Calculated from data.

c. Lilliefors Significance Correction.

d. This is a lower bound of the true significance.

The hypotheses comparing the mean of video group’s number of actions with 14 reference actions are as follows:

Null hypothesis (H0): The mean of found actions is equal to 14.

Alternative hypothesis (H1): The mean of found actions is not equal to 14.

According to Table 6.49, the mean of determined actions is 9.30 which is not equal to 14, but its significance should be tested. Table 6.49 demonstrates the t-test results including the value of t, 2-tailed significant level, and confidence interval of 95%.

Table 6.49: Mean comparison test results in video group

One-Sample Test						
	Test Value = 14					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Number of actions	-5.569	9	.000	-4.700	-6.61	-2.79

The 2-tailed p-value of zero which is less than 0.025 shows that the null hypothesis is rejected, and the mean of determined actions (9.3) is not equal to the reference action (14), and there is a significant difference between the two.

The same study applies to the written narration group. There are no outliers in the boxplot, and although the median line is not in the middle of the boxplot, the boxplot is not skewed (Figure 6.38). Therefore, the populations appears to be normal. To confirm the boxplot results, the Kolmogorov-Smirnov test is used, and the results are as follows:

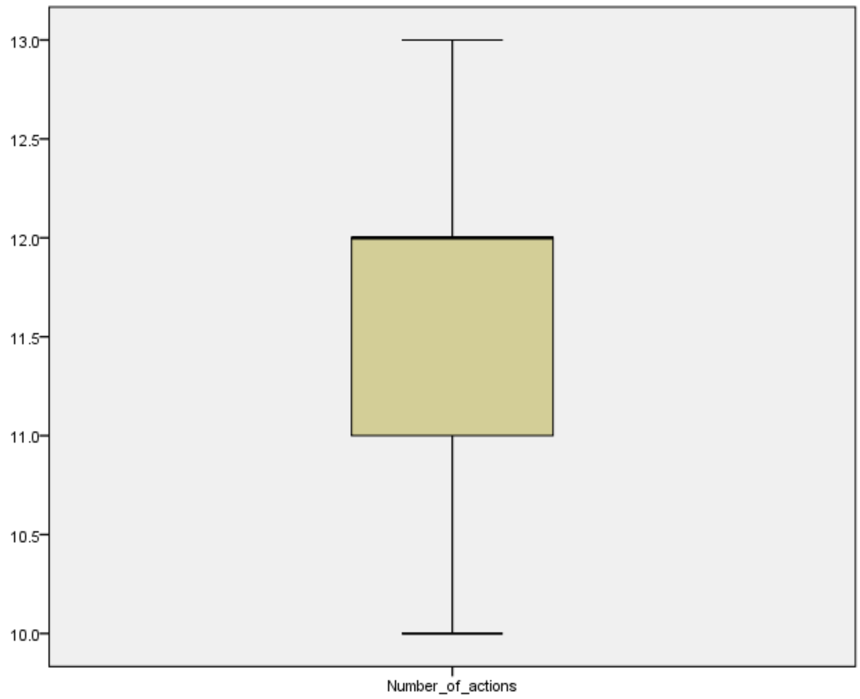
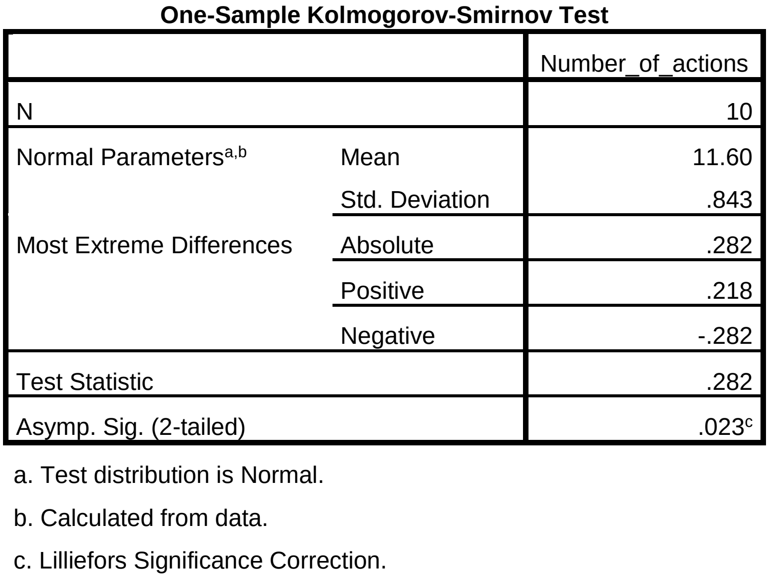


Figure 6.38: Boxplot of number of actions in written narration group

The Kolmogorov-Smirnov test shows (Table 6.50) that the population is normal.

Table 6.50: Kolmogorov-Smirnov test results for written narration group



One-Sample Kolmogorov-Smirnov Test

		Number_of_actions
N		10
Normal Parameters[a,b]	Mean	11.60
	Std. Deviation	.843
Most Extreme Differences	Absolute	.282
	Positive	.218
	Negative	-.282
Test Statistic		.282
Asymp. Sig. (2-tailed)		.023[c]

a. Test distribution is Normal.

b. Calculated from data.

c. Lilliefors Significance Correction.

The hypotheses comparing the mean of written narration group's number of actions with 11 reference actions are as follows:

Null hypothesis (H0): The mean of found actions is equal to 11.

Alternative hypothesis (H1): The mean of found actions is not equal to 11.

According to Table 6.51, the mean of found actions is 11.60 (Table 6.51) which is not equal, but very close, to 11. The significance of difference is tested, and the t-test results are shown in Table 6.52.

Table 6.51: Descriptive statistics of number of actions in written narration group

One-Sample Statistics				
Variables	N	Mean	Std. Deviation	Std. Error Mean
Number of actions	10	11.60	.843	.267

Table 6.52: Mean comparison of test results in written narration group

One-Sample Test						
Variables	Test Value = 11					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Number of actions	2.250	9	.051	.600	.00	1.20

The 2-tailed p-value of 0.051 which is greater than 0.025 and descriptive statistics of 11.60, show that the null hypothesis is accepted, and the mean of found actions is not significantly different from the 11 reference actions.

6.4.2. Coded Actions Analysis

After checking the average number of actions determined by participants in both groups, the differences between their answers is examined in the following. The different categories of answers are Correct Actions, Missing Actions, and Wrong Actions.

As shown in Tables 6.52 and 6.53, a number of coded actions that were irrelevant or additional to the scenario were omitted from the number of coded actions. The number of missing actions in the videotaped scenario group is higher than for the written scenario group. The comparison between the rate of the missing actions of the both groups shows that the missing actions were caused by human error. The number of the wrong actions is very low in both groups, thereby indicating the applicability of the taxonomy.

Table 6.53: Videotaped scenario group data

User ID	Number of coded actions	Correct actions	Missing actions	Wrong actions	Number of Accepted Actions	Total
1	9	7	6	1	8	14
2	7	6	8	0	6	14
3	13	11	2	1	12	14
4	11	9	4	1	10	14
5	11	9	3	2	11	14
6	7	6	7	1	7	14
7	9	6	5	3	9	14
8	8	7	6	1	8	14
9	5	4	9	1	5	14
10	13	11	2	1	12	14

Table 6.54: Written scenario group data

User ID	Number of coded actions	Correct actions	Missing actions	Wrong actions	Number of Accepted Actions	Total
1	10	10	1	0	10	11
2	11	11	0	0	11	11
3	12	9	0	2	11	11
4	11	11	0	0	11	11
5	11	11	0	0	11	11
6	13	7	3	1	8	11
7	12	10	0	1	11	11
8	12	10	0	1	11	11
9	12	9	1	1	10	11
10	12	11	0	0	11	11

Figure 6.39 illustrates the inverse relationship between the missing and the correct actions in the videotapes scenario group. The results for the correct actions can be clustered into three groups: high (≥ 60%), medium (≥ 45 and ≤ 60%), and low (≤ 45%).

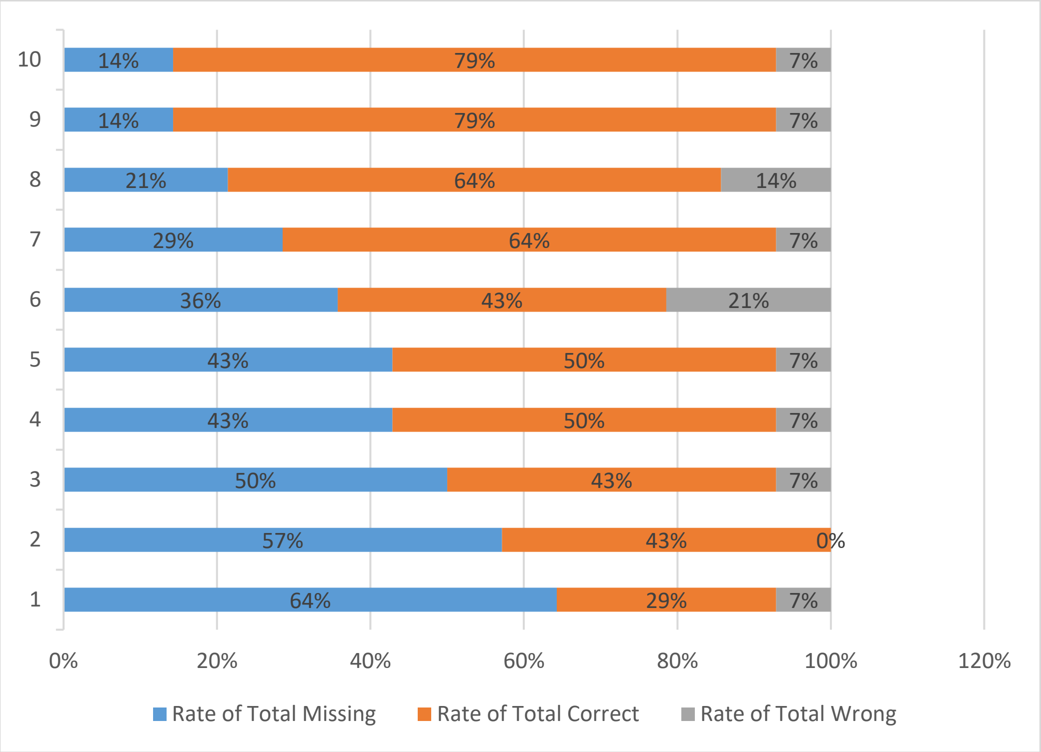


Figure 6.39: The Videotaped Scenario Group Results

Based on these three groups, all the participants in the written scenario group (Figure 6.40) are ranked highly, and 60% of the videotaped scenario group are in the medium and high groups. Those participants with the low results missed the highest number of actions.

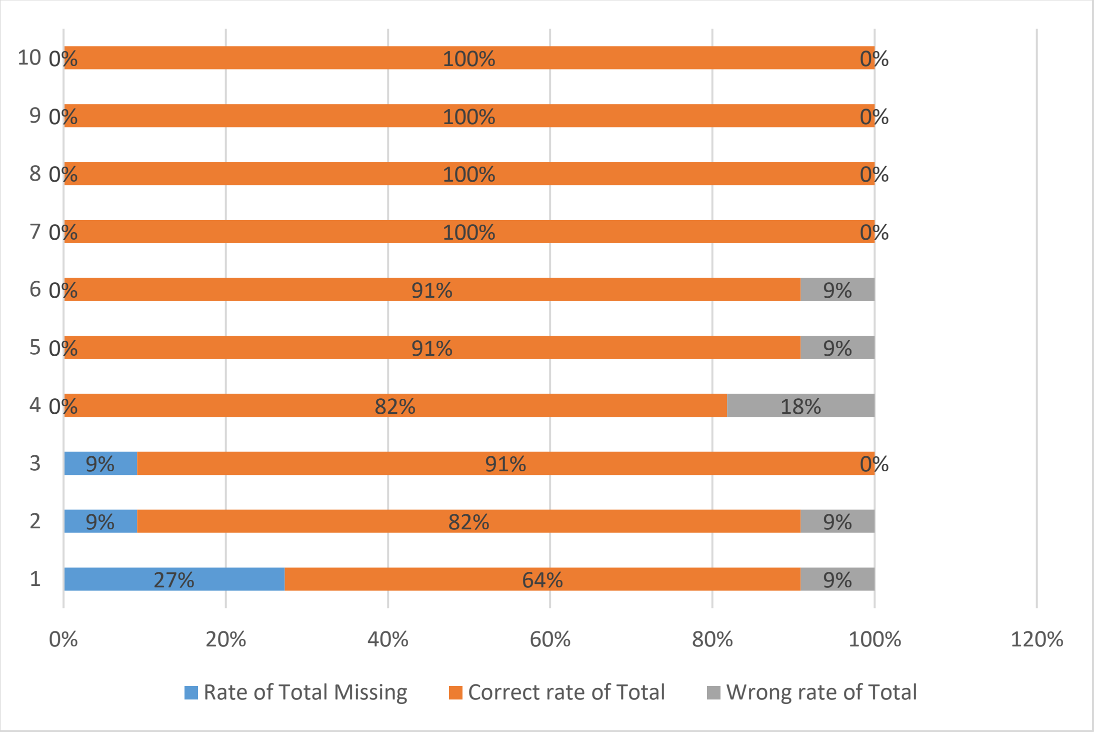


Figure 6.40: The Written Scenario Group Results

6.4.3. Participants' Feedback Analysis

After finishing the coding task, participants were asked to provide feedback on the use of the taxonomy, actions and their attributes, codified actions, their exhaustiveness, and the similarity of the codified performance with observed real-life performance. Participants described their experience with codifying the given scenario by using BEHAVE's action classification and attributes.

Responses regarding the taxonomy and its use fall into two categories:

1. 'Confirming the power of codification',
2. 'Being not sure completely about the power of codification'.

To reflect their thoughts, participants in the first category used phrases such as:

- "The way it is coded is very well suited";
- "I didn't have that much problems to code the sentences";
- "I think it's quite comprehensive";
- "it does make sense even if you are missing some words like filling words";

- "I think this fit quite well";
- "you just tick off what they are doing";
- "I don’t think there was any other issues";
- "it should be enough";
- "they were sufficient";
- "they were clearly defined and it was easy to identify".

A few participants, who were in the category ‘Being not sure completely about the power of codification’, expressed their opinion with phrases such as:

- "you need to give somehow more protocols to people who want to convert the actions to codes";
- "they were sufficient";
- "they weren't particularly precise but for classifying (what I think is their purpose) they seemed to perform";
- "I found I was able to code most of the actions";
- "I think the number and type of the actions depends on the case".

When commenting on the similarity between codification and observation of real-life performance, participants used the following phrases in their feedback:

- "I can understand both in the same way";
- "the outcome of both would be similar to me compare to real life";
- "if I want to quantify I would say like 90%";
- "I think it’s pretty close";
- "pretty close I guess";
- "you have all the information without all the extra words that are not necessary";
- "I think pretty close";
- "they dealt with the situations sufficiently";

- "I think the classes of actions and the attributes are developed well and can simulate a real life situation properly";
- "I think the coding, for the most part, was really close to the video";
- "I could code the performance in 90% compared to the real life situation";
- "it was enough attributes in the code to get close to the real situation".

Eighty percent of participants indicated that the classes of actions and the different attributes provided to describe those actions were sufficient to code the observed performance; 20% of the participants stated that there might be room for improvement in terms of more protocols and re-arranging of the attributes.

Responding to the similarity question, 70% of the participants strongly suggested that the codification is very similar to observed real-life performance. The remaining 30% were positive that the codification can cover more than 90% of the observed performance (Table 6.55).

Table 6.55: Concept frequency

Concepts	**Frequency**
They were sufficient to code everything	80%
Almost enough to code everything	20%
Strongly suggest that codification is close to real-life performance	70%
Being positive (90% and above) that codification is close to real-life performance	30%

Figure 6.41 projects the relative frequencies of the highest ranked descriptive adjectives and words used by participants, which shows a high use rate of positive adjectives such as ‘good’, ‘close’, ‘clear’, and ‘comprehensive’.

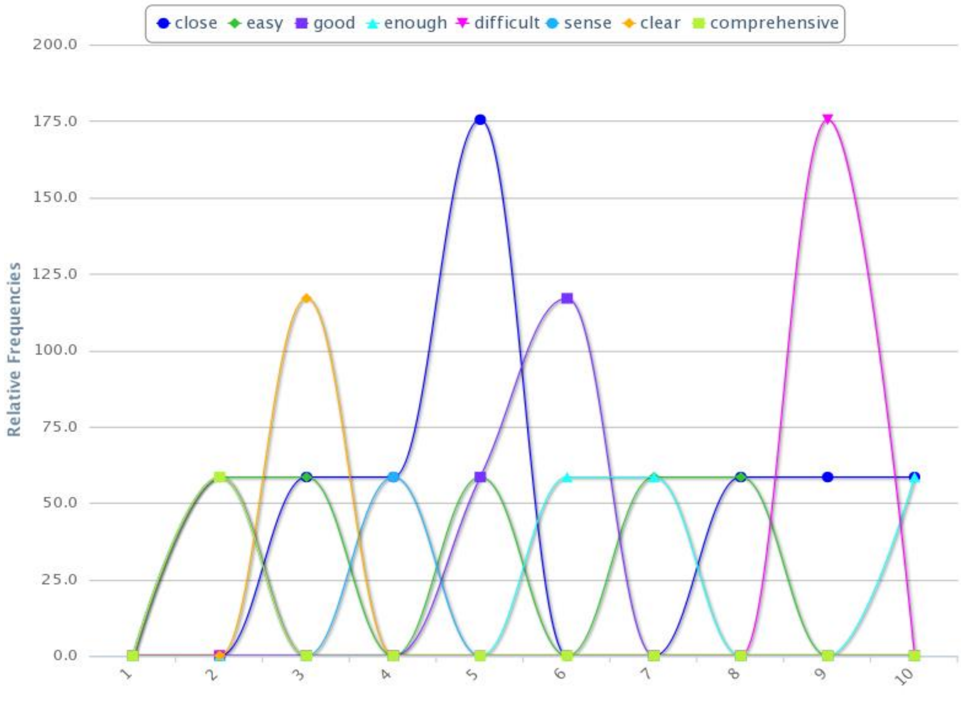


Figure 6.41: Relative frequencies for descriptive words used by participants

Chapter 7: Discussions

7.1. Introduction

In this chapter, the results and analysis presented in Chapter 6 are discussed with respect to the validity of BEHAVE; this is followed by a discussion of the evaluation of the taxonomy. The chapter concludes with a summary of the discussions.

7.2. Survey for Expert Opinion: VW and Industry Experts

Expert opinion is an evaluation method used in both DSR and taxonomy evaluation. The survey was used during the development of BEHAVE to help determine the face validity of the taxonomy.

As explained in Chapter 5, the BEHAVE levels were inspired by Action-based Learning Assessment Methodology while the classes of Functional Acts were established after an extensive literature review of taxonomies and theories of human actions and direct observation of human actions in both real and simulated environments, complemented by the researcher’s expertise and knowledge. Having said that, the expert opinion survey contributed significantly to the development of BEHAVE.

The first section of the survey (Questions 1, 2, and 3) investigated the importance of the three levels of actions in BEHAVE. Considering the high degree of importance given by the experts, the study and comparison of the experts’ opinion in both groups shows that the Goal Act is unanimously ranked as ‘very important’. As discussed in Section 4.2, Action-based Learning assessment scenarios, specifically ALAM, require learners to achieve an ‘Outcome goal’ that may be a problem to solve, a scenario to be completed, or other possible expectations to be fulfilled. This is aligned with the results of the survey: the Goal Act ranked as the most important taxonomy level.

It is noted that the experts in the VW group ranked the Constitutive Acts higher than the Functional Acts while the industry experts gave a higher ranking to the Functional Acts. However, the lack of unanimous opinion does not seem to undermine the importance of these two levels of the taxonomy (due to the highly ranked importance). The reason for the higher priority given to the Functional Acts by the industry experts reflects the importance of the atomic actions that constitute performance in real-life working situations. This priority can be explained by the fact that real-life performance directly affects the health and safety of the person, while even a death in VW will not threaten anyone's safety.

Moreover, the experts were asked to rank the importance of the BEHAVE classes of actions. Although the experts were not unanimous in their opinion, the industry experts ranked the taxonomy classes from 'important' to 'very important', while the VW experts ranked them 'important'. This shows that experts from industry have a more consensual opinion about the importance of different classes of human actions rather than recognising only the Goal Act as the most important level. These results are aligned with the results of ranking the levels of actions in the first section of the survey, where the industry experts gave a higher priority to Functional Acts. These results also indicate the importance of the ALAM's support for authentic assessment, given the important role of the Functional Acts in real-life settings.

Furthermore, the importance of multiple expert reference solutions is noted in the analysis of the survey results. Sixty-one percent of the experts believed that it is better to compare the learners' actions with 1-3 reference solutions, and 31% believed this number should be between 3 and 5. The 92% agreement in multiple reference solutions by experts supports the multiple reference solution characteristics in ALAM.

A comparison of the highly ranked importance of the need for a taxonomy of human actions in Action-based Learning assessment with the low number of experts naming any taxonomies of human actions they might know, indicated a gap in the extant research and motivated the development of BEHAVE.

Finally, the 'Chi-Square Test of Homogeneity' results demonstrate that, despite the fact that the respondents ranked highly the importance of the questions, there is a balanced difference of opinion among the experts, both inter-groups and intra-groups. This balanced difference of opinions and a high Cronbach's Alpha reliability coefficient, demonstrates a high degree of reliability of the contribution of the survey as an expert guidance and feedback in the research process.

7.3. Card Sorting Test

In this research, the card sorting test was used to examine the exclusiveness of the BEHAVE classes. The results were analysed using the Scatterplot matrix, K-mean Cluster analysis, Pearson correlation test, Fowlkes and Mallows index, and R-squared (Section 6.3) in order to determine the different aspects of the internal validity of the taxonomy.

The results from the card sorting test were analysed using the K-mean cluster analysis method (Section 6.3.1) which indicated six separate clusters. As the initial number of clusters used in both the test and K-mean cluster analyses was the same, the K-mean cluster analysis was conducted using five and seven clusters as well (Figure 6.35) to investigate the possibility of any other acceptable cluster numbers. The three dissimilar number of clusters in the K-mean cluster analysis indicates that the initial number of the chosen classes was the best choice since, by reducing or increasing the number of the classes, there would be a class with a high level of overlapping of Operative and Constructional actions.

The resulting Fowlkes and Mallows index (Section 6.1.2) calculated for the clusters (0.95326) shows a lengthy distance between the clusters, indicating the exclusiveness of the clusters that represent the classes of Functional Acts in BEHAVE. The substantial distance between the cluster centres is also supported by comparing the results in Tables 6.39 and 6.40. The results show that the sum of maximum distances from the cluster centre for each pair is less than the distance between each pair of cluster centres. As a result, it is evident that there is no overlapping between the clusters.

To demonstrate the exclusiveness of the classes, the clusters were also tested using the Pearson correlation test that indicates ‘Weak’ or ‘Very Weak’ correlations between the clusters (Table 7.1). The very weak or weak correlation between the clusters supports the results of the K-mean cluster analysis.

Table 7.1: Correlations between the BEHAVE classes based on Pearson correlation test results interpreted by Evans & Over (2013) suggestions.

	Correlations					
	Constructional	Decisional	Gestural	Locomotive	Operative	Responsive
Constructional	1	Weak	Weak	Very Weak	Very Weak	Weak
Decisional		1	Weak	Weak	Weak	Very Weak
Gestural			1	Very Weak	Weak	Weak
Locomotive				1	Very Weak	Weak
Operative					1	Very Weak
Responsive						1

However, the scatter plot matrix (Figure 6.34) shows a number of actions classified with a noticeable distance from the regression line. As the histograms, shown on the diagonal of the scatter plot matrix, indicate, there is a clear distinction between the actions classified under the main cluster and the rest of the data. This is aligned with the resulted R-square ($R^2 = 0.95326$), indicating a low dispersion of data in each cluster. It also supports the discussion on the barriers of language and context (Sections 5.4 and 5.5) in taxonomies of human actions and performances.

The exclusiveness of the BEHAVE classes of Functional Acts is strongly demonstrated by the appropriate number of clusters, the correlation between the clusters, the distance between the cluster centres, and the dispersion of data in each cluster.

7.4. Performance Coding Experiment

7.4.1. Participants' Coding

The results of this experiment reveal that there is a meaningful difference between the number of actions determined by participants in the video group and those determined by the participants in the written narration group. While the written narration group had a very close mean of determined actions to the number of reference actions (11.60 vs. 11), there is a big difference in the video group (9.30 vs. 14). The results suggest an increase in the number of actions determined by the participants with a decrease in the level of freedom. Furthermore, the results of the coded actions by both groups show that the video group participants have the most missing actions, while the written narration group shows the highest number of correctly identified actions compared to the video group. This also confirms that by increasing the freedom in judgment and decreasing the structured information, the human participants are more likely to make errors such as missing performed actions. In other words, the more the participant is constrained by structures in the representation of performance, the fewer errors are made when determining the performed actions (written scenario vs. videotaped performance).

In other recognised actions, such as ‘correct but different actions’, ‘partial actions’, ‘irrelevant actions’, ‘wrong actions’, ‘correct additional actions’, and ‘correct misplaced actions’, there was no significant difference between the two groups. An important point in these results is the low number of actions in both groups, and their slight difference. This indicates that the taxonomy was successful in creating a standard procedure for classifying and describing performed actions, as the other types of actions are correct even if they are misplaced, additional, or partial; this is also an indication of human error.

As explained, the major difference between the two groups of participants is in the number of determined, missing, and correct actions; this is regardless of the use of the taxonomy or the coding tool. The participants in the group with the written scenario had the average number of determined actions close to but higher than the reference actions (indicating that the participants added extra actions based on their interpretation or prior knowledge of the field), while the participants in the videotaped scenario group identified significantly fewer performed actions in the videotaped performance. The low rate of correctly identified actions by the videotape group is due to the high rate of missing actions; however, the group with the written material had a very low level of missing and wrong actions. It is evident that errors such as missing the actions, partially coding the actions, and misplacing the actions, are human errors that were significantly increased due to the increase in the degree of freedom in judgement and interpretation.

Overall, the experiment results and analysis strongly suggest the use of a computerised representation of the performed actions, during both the assessment and analysis. As can be seen, human examiners are prone to making a variety of human errors. Human examiners may have different interpretations of the same performance, and the outcome of the assessment, including the feedback from the examiner, might not be consistent. (2D/3D) VTE as the assessment environment in ALAM, enables a standard and consistent digital representation of human actions. BEHAVE provides a formal classification and syntax to code those actions; consequently, ALAM uses actions coded consistently with the same syntax (for both learner and expert) to compare and generate a formative feedback.

This experiment was designed to evaluate the taxonomy for external validity by showing its usability and applicability. The main goal of this experiment was to determine the degree of precision with which actions can be classified, described, and coded by using the BEHAVE classification, attributes and syntax. The results show a high degree of precision for most of the correct actions and low rate of wrong actions, which proves the validity of the taxonomy.

In the following section, the feedback from the participants is investigated to discover the degree of similarity of the generated codes to the real-life performance and the exhaustiveness of the taxonomy according to the participants.

7.4.2. Participants' Feedback

From the participants' feedback, the researcher noted that it was necessary to improve the BEHAVE syntax. Hence, following a participant's suggestion, improvements were made after consultation with a linguistics expert who advised that the attributes in the syntax be rearranged.

The feedback provided by the participants who were not certain about the use of the taxonomy indicated that if the taxonomy is to be used by humans, there should be clear guidelines about the appropriate method of classifying and coding the actions. Although the feedback from participants showed some uncertainty, it nevertheless indicated a positive attitude towards the taxonomy.

The participants' responses indicated that most of them believed that the given classes of actions and different attributes could cover almost all of the recognised actions. Participants stated that they could describe the actions very closely to the real-life occurrences by using the given attributes.

7.5. Taxonomy Evaluation

The use of DSR as the research methodology provided an opportunity to redefine the solution to the research problem being investigated. The constant evaluation of the solution provided clear results on the internal and external validity of BEHAVE. The validity tests used in the evaluation phase of the research addressed the criteria that need to be satisfied for the internal and external validity of taxonomy. These criteria for internal validity are mutually exclusive classes on the horizontal level, and exhaustiveness. For external validity, the degree of adaptability, usability, and also the practical applicability of the taxonomy in different fields of study are the most important criteria. External validity is mostly tested by means of experiments.

Two important criteria for the internal validation of taxonomy as stated by Fleishman et al. (1984) are: having 'mutually exclusive classes on the horizontal level' which put each entity under just one class, and the second is being 'exhaustive', which enables every entity to fall under a class one way or another. As Fleishman et al. (1984) explained, performance and human action taxonomies have the most difficulty meeting the first criterion, 'mutually exclusive classes on the horizontal level'. In this research, multiple approaches were taken to evaluate and prove the internal validity of the BEHAVE taxonomy of human actions, including a card sorting internal validity test, an experiment, and application to diverse scenarios.

As can be seen in Chapter 5, Section 5.3, the BEHAVE syntax for action codification has been applied to different scenarios including an HTA example, to demonstrate the versatility of the performances that can be coded by BEHAVE. However, it cannot be claimed that using BEHAVE to code different scenarios demonstrates the exhaustiveness or generalisation of the taxonomy. However, a variety of scenarios in dissimilar fields can be coded by BEHAVE, thereby demonstrating the range of actions that BEHAVE covers. The number and range of scenarios and actions also helps to confirm the external validity by showing the adaptability of the taxonomy.

Similarly, in an experiment conducted in this research, feedback gathered from the participants clearly indicated that there was no human action for which the participants could not find a suitable class under the Functional Acts classes (Section 7.4.2). BEHAVE showed a high degree of exhaustiveness in both examples and participants' feedback.

Furthermore, the cluster analysis results (Section 6.3) of the card sorting test demonstrate a high level of satisfaction with the first criterion, mutually exclusive classes on the horizontal level. The results show that the actions are similarly clustered by participants, by the researcher, and by the cluster analysis method. They also show a 95% (Section 6.3.3) internal validity for the cluster analysis, which is considered a high degree of validity for cluster analysis. The cluster analysis results satisfy the first criterion of internal validity, the ‘mutually exclusive classes on the horizontal level’.

However, one might raise the issue that the descriptive diagram in Figure 6.36 shows that some actions such as ‘run’ are placed under different classes. As discussed above, it is difficult for human action taxonomies to meet the first internal validity criterion of exclusiveness, although the BEHAVE cluster analysis shows a high degree of validity. This issue is addressed in Chapter 5, Sections 5.4 and 3.5, which discusses context and language barriers. Various suggestions are made to overcome these barriers. Nevertheless, the statistical analysis proves the internal validity of the six clusters or classes of actions.

The external validity should be evaluated by means of an experiment or lab test. In this research, a performance coding experiment was designed and conducted to evaluate the external validity and applicability of the taxonomy. Results clearly show that by imposing greater restriction on human interpretation, there is a reduction of errors and also an increase in the number of correctly recognised actions. The written narration of a scenario clearly produced better results and a higher rate of correctly codified actions than did the videotaped performance.

Considering that the video recording is the closest reflection of the real-life performance, the rate of correctly coded actions among participants watching the video was significantly different from that of the participants who were given the written narration. The descriptive data and codes generated by these experts show the effect of diverse interpretations that may decrease the level of standardisation in assessment. The interesting aspect of the experiment results was that the number of incorrect actions was very low in both groups, and most mistakes were made by missing the action, adding extra actions, or misplacing the actions.

All the participants stated in their feedback that they could easily recognise the class of the actions and their attributes, and it was very close to the real-life situation. This shows that the difference between the participants’ feedback and their results are caused by human errors such as missing actions or misplacing them. Therefore, it can be concluded that BEHAVE was successful in being usable and applicable regardless of the human errors that caused a significant reduction in the number of actions that were coded correctly both in description and sequence. Hence, the evaluation results indicate that the taxonomy has both internal and external validity.

7.6. Summary

To summarise, in this chapter the contributions of the experts opinions gathered via a survey were discussed. Moreover, the exclusiveness and exhaustiveness of the BEHAVE classes were discussed using the card sorting test results ‘presented in Chapter 6. Furthermore, the results of the coding experiment and the feedback from the participants in this experiment were discussed. Finally, the internal and external validity of BEHAVE as a taxonomy of human actions and DSR artefact was investigated.

From the discussion of results, it is important to note that: BEHAVE is demonstrably a valid taxonomy of human action; BEHAVE enables ALAM to create consistency in its comparison and automated formative feedback generation; although BEHAVE is designed to be used by both humans and computers, the use of BEHAVE in a computerised setting enables BEHAVE to fulfil its purpose of classifying and describing the performed human actions in any environment, especially a computerised environment, with a high degree of similarity through its classification of human actions, action-attributes set, and formalised syntax.

The next chapter reviews the research gap that motivated this research; demonstrates how the research objectives have been achieved and the research question answered; and sets the stage for future research.

Chapter 8: The Past, Present and Future

8.1. Introduction

‘Past, Present, and the Future’ summarises how this research addressed the recognised research gap and set the stage for the future research. In the ‘Past’ section, the research gap is stated, and the section ‘Present’ demonstrates how the research objectives have been achieved, thereby answering the research question. The practicality and constraints of the research are also discussed under ‘Present’. This chapter also discusses the various fields of research that can benefit from the findings of this research and suggests several future research opportunities. The chapter concludes with a summary of the thesis.

8.2. Past

An investigation of the various learning and assessment theories and technologies showed a high demand for different experience-based methods (Section 2.1) with more opportunities of authentic and immersive learning and assessment environments (Sections 2.2.3, 2.3.3, and 2.4). The literature indicated a need for an assessment methodology that automatically generates formative feedback based on learners’ actions during the assessment process.

The literature review (Chapter 2, Section 2.5), revealed that different researchers have developed various classifications of human actions with diverse applications. These studies included task analysis, performance analysis, human error analysis, gestures recognition, video recognition, and computer-supported cooperative work. An examination of human action theories showed that most of the developed taxonomies are specifically focused without any intention of having additional extensions and adaptations. Thus, there was an evident need for a taxonomy which is flexible, allows adaptation and transfer to other scenarios and contexts, and is not constrained to the original application.

8.3. Present

8.3.1. Research Question and Objectives

This research was designed to respond to the research question: How can learners' actions be formally represented to create consistency in the assessment process leading to an automated post-performance formative feedback? To answer the research question, the researcher needed to achieve the following objectives by means of the DSR methodology:

1. To analyse the literature of taxonomies and classifications of human actions in different disciplines (Section 2.5).

As summarised in Table 2.6, human actions have been the focus of several theories, classifications, and taxonomies such as actions, activities, tasks, and performances in real-life or simulated settings. These taxonomies and classifications were studied extensively during this research with the purpose of finding a taxonomy of human actions to be used in PCE component of ALAM, or to provide the foundation for a new taxonomy.

As discussed previously and as demonstrated in Table 2.6, the studied taxonomies and classifications were too field-specific or too general for use. Nonetheless, each taxonomy had its unique attributes that could be used in the development of an exhaustive taxonomy of human actions. Exhaustiveness in this context would be a taxonomy that is not field-specific or general, but covers a wide range of actions regardless of their field. The feedback results from the experiment conducted in this research (Section 6.4.3) demonstrated a high degree of satisfaction among the participants concerning the exhaustiveness of the taxonomy classes in covering all the actions considered by the participants.

Although the taxonomies summarised in Table 2.6 contributed to this research in part, the following theories and taxonomies constituted the main contribution to the formation of BEHAVE:

- The theory of human actions (Goldman, 1970);
- Taxonomy of Embodied Actions for cooperative design in a distributed company (Robertson, 1997, 2000);
- Avatar Capabilities Model (Chodos et al., 2014).

2. To develop a classification of human actions (Section 5.2).

As explained in Section 5.2, BEHAVE was developed using primary and secondary research, collecting data by means of an expert opinion survey, and literature reviews. However, the expertise of the researcher played a substantial role in the choice of a taxonomy development method. The secondary research, with the data extracted from the literature, was used as a foundation for BEHAVE. Following an intensive study of the literature and analysis of the previously mentioned theories and taxonomies, the researcher used his expertise to develop the BEHAVE levels, classes, and the action coding syntax. Direct observation of human actions in both real and simulated environments played a key role in the development of the BEHAVE classes. The primary data collected via the survey was also used for development and evaluation purposes.

Cluster analysis of the card sorting data (Section 6.3) showed exclusive classes in the taxonomy, demonstrating the validity of the developed classes. Moreover, the expert opinion survey demonstrated that experts ranked the BEHAVE levels as ‘important’ to ‘very important’ (Section 6.2).

BEHAVE classifies human actions in three main levels and six classes. The levels are the Goal Act, Constitutional Acts, and Functional Acts. The Goal Act consists of one or more Constitutive Acts, which include one or more Functional Acts. Functional Acts, as the most basic action level, are classified into six different classes: Gestural, Responsive, Decisional, Operative, Constructional, and Locomotive.

3. To develop a set of action attributes to describe the actions (Section 5.2.2).

BEHAVE uses a set of action attributes to describe each action. The attributes are Preposition, Adjective, Object, Quantity, Unit, Property, and Location. Preposition is used to show the relationship in terms of space and time. The Object is the thing at which the performed action is directed. Although the Object is targeted by the performed action, it might not necessarily be altered. The Adjective is used to qualify the Object, Location, or the performed action. The Property, Quantity and Unit are used to describe a physical property such as diameter or temperature. The Location can be used to specify the place in which the performance is occurring; this place can be the main venue such as kitchen, shop or lab; or it can refer to the place where the object of the action is placed, such as machine, pot, or basin (Section 3.2.2). According to the feedback results from the experiment conducted in this research (Section 6.4.3), the participants indicated that the attributes were sufficient to describe the performed action in the coding process.

4. To design a formal syntax to structure the actions as computer-readable data (Section 5.2.2).

BEHAVE uses a syntax to structure the performed actions (Section 5.2.2). The syntax is as follows:

[<Action.Levels>]<Trigger.Action><Action.Class><Action.Type>[Preposition, Adjective, Object, Quantity, Unit, Property, Location][Rules][Timestamp]

Attributes are applied to each action in order to describe the action similar to the real-life performed action. The Rules are adopted from logical relationships (dependency rules) in project management (Project Management Institute, 2011) including Finish-to-Finish (FF); Finish-to-Start (FS); Start-to-Finish (SF); and Start-to-Start (SS). The Timestamp establishes the sequential order of the performed actions. The timing of actions can be used to recognise behaviours such as hesitation or uncertainty. This syntax was used and tested in an experiment of similarity degree of the coded scenario compared to real-life performance (Section 6.4).

5. To evaluate the taxonomy for internal and external validity (Section 7.5).

BEHAVE was evaluated as a DSR artefact in order to determine its internal and external validity, using an experts' opinion survey, card sorting, performance coding experiment, and participant feedback (Chapter 6). The survey provided experts' opinions on the importance of BEHAVE levels and classes, confirmed by cluster analysis. The clusters resulted from the card sorting test, demonstrating the exclusiveness of BEHAVE classes, and consequently, its internal validity. The performance coding experiment was conducted to examine the usability and applicability of the taxonomy according to the feedback provided by the participants in the experiment. The analysis of the experiment results and feedbacks indicated the external validity of BEHAVE.

8.3.2. Research Practicality

As shown by the conducted experiment (Section 6.4), card sorting test (Section 6.3), and application to different scenarios (Sections 5.3), the taxonomy classes cover a widespread range of human actions, and the action-attributes set describes these actions adequately close to real life. BEHAVE can be employed in different disciplines to standardise the outputs, which make them usable cross platform.

The syntax enables BEHAVE to be utilised in various computer technologies such as artificial intelligence. Although BEHAVE has been developed for computer-mediated action classification and codification, it can easily be used manually by human experts (Section 6.4).

8.3.3. What Are The Constraints?

This research, similar to every research involving humans, was subject to certain constraints including limitations of time, finance, human participants and, possibly, computer programming. The availability of data regarding performed actions in current (2D/3D) VTEs was limited, and acquiring the appropriate data required extensive programming which was not feasible given the constraints of time and finance. As the assessment system based on the ALAM framework was in its conceptual development stage, BEHAVE could not be evaluated by being used by the system, which provides a promising opportunity for future research.

Although this research faced these diverse constraints, none of these limitations compromised the quality of the conducted research. The time and financial limitations of a Ph.D. research were considered in the research design and were overcome by the precise research focus and accurate implementation of the research design and DSR steps. The constraints regarding human participants were dealt with by careful choice of sampling, data gathering, and data analysis methods. The lack of computerised data input and processing was addressed by means of manual reconstructions.

8.4. Future

During the secondary literature review, the potential for various applications of BEHAVE in different disciplines such as human reliability applications, video recognition, error recognition, pattern recognition, and artificial intelligence was presented. Although these technologies and their applications are independent, they can contribute to other solutions as interconnected components. This interoperability requires a standard method of communication that can be provided by BEHAVE. Furthermore, BEHAVE can be useful in human factors research in, for example, the fields of behavioural science, applied psychology, ergonomics, and engineering.

Action-based assessment in 3D VTEs will use BEHAVE for a formalised representation of performed actions. None of the current 3D VTEs, and on a larger scale VWs, are designed based on human actions; however, there is a great opportunity to add this ability to some of the current VTEs. Therefore, these environments are not able to provide descriptive data regarding the performed actions. The lack of a standard data set for the purpose of learning assessment provides an ideal opportunity for development of a standard action set for 3D VTEs and VWs. Depending on their different capabilities, artificial intelligent systems can also be used for consequence prediction in a variety of situations.

8.5. Conclusion

In this thesis, the author presented the research leading to the creation of a taxonomy of human actions, BEHAVE, for Action-based Learning Assessment in 3D Virtual Training Environments. The following recapitulates the chapter contents.

Introduction: introduces the research and its motivation, purpose, and significances.

Literature review: investigates the literature pertaining to Action-based Learning methodology and its various theories and methods, assessment and its different types, 3D virtual environments and their application in learning and assessment, and the various taxonomies of human actions, tasks, and behaviours.

Research methodology: includes the research aim and objectives, the research design, and the various data gathering and data analysis tools used in this research.

ALAM: introduces the ALAM framework and conceptual model for a software system based on ALAM.

BEHAVE: provides a exhaustive explanation of BEHAVE taxonomy of human actions including the definitions, classifications, examples, syntax and rules, and contextual and lingual concerns.

Evaluation results: presents statistical results of experiments and tests performed in the evaluation phase of the research including survey, card sorting, and experimentation.

Data analysis and interpretation: examines the results and data analysis and interpretations.

Past, present, and future: the thesis concludes with a summary of the findings and suggestions for future research undertakings.

The chapters acquaint the reader with: the desirability of using BEHAVE to generate formative feedback in an automated action-based assessment method; the definition and structure of BEHAVE; and the evaluation of BEHAVE for internal and external validity.

The various validity tests performed during this research (Chapters 6 and 7) indicate that BEHAVE is a valid taxonomy both internally and externally. The validity of the taxonomy, and its flexibility, make it appropriate for use in different fields of study as a common taxonomy of human actions.

REFERENCES

Abrams, S. S. & Gerber, H. R. (2013). Achieving through the feedback loop: Videogames, authentic assessment and meaningful learning. *English Journal*, 103(1), 95-103.

Aggarwal, J. K. & Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Computer Survey*, 43(3), 1-43.

Aken, J. E. V. (2004). Management research based on the paradigm of the design sciences: the quest for field‐tested and grounded technological rules. *Journal of management studies*, 41(2), 219-246.

Allen, J. F. (1984). Towards a general theory of action and time. *Artificial Intelligence*, 23(2), 123 - 154.

AL-Smadi, M., Gutl, C., & Kannan, R. (2010). Article: Modular Assessment System for Modern Learning Settings: MASS. *International Journal of Computer Applications*, 1(9), 43–49.

Alturki, A., Gable, G. G., & Bandara, W. (2011). A design science research roadmap. In *Service-Oriented Perspectives in Design Science Research* (pp. 107-123). Springer Berlin Heidelberg.

Annett, J. (2004). Hierarchical Task Analysis (HTA). In N. A. Stanton, A. Hedge, K. Brookhuis, E. Salas & H. W. Hendrick (Eds.), *Handbook of Human Factors and Ergonomics Methods* (Vols. 1–0, pp. 33–1–33–7). Boca Raton, FL, USA: CRC Press.

Anscombe, G. E. M. (1958). *Intention*. B. Blackwell.

Ashford-Rowe, K., Herrington, J. & Brown, C. (2014). Establishing the critical elements that determine authentic assessment. *Assessment & Evaluation in Higher Education*, 39(2), 205-222.

Baehr, M. (2005). *Distinctions between Assessment and Evaluation, module in 2nd edition of the Faculty Guidebook*, Lisle, IL: Pacific Crest.

Bailey, K. D. (Ed.). (1994). *Typologies and Taxonomies*. Thousand Oaks, CA, USA: SAGE Publications, Inc.

Barab, S. A., Squire, K. D., & Dueber, W. (2000). A co-evolutionary model for supporting the emergence of authenticity. *Educational technology research and development*, 48(2), 37-62.

Barrows, H. S. & Tamblyn, R. M. (1980). *Problem-based learning: An approach to medical education.* New York, NY, USA: Springer Publishing Company.

Barrows, H. S. (1996). Problem-based learning in medicine and beyond: A brief overview. In L. Wilkerson & W. H. Gijselaers (Eds.), *Bring problem-based* *learning to higher education: Theory and practice*, Vol. 68. (pp. 3–12). San Franscisco, CA, USA: Jossey-Bass.

Beard, C., & Wilson, J. P. (2013). *Experiential learning: A handbook for education, training and coaching* (p. 343). London, UK: Kogan Page Publishers.

Bell, M. W. (2008). Toward a definition of "virtual worlds". *Journal for Virtual Worlds Research*, 1(1).

Belland, B. (2012). The Role of Construct Definition in the Creation of Formative Assessments in Game-Based Learning. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in Game-Based Learning* (pp. 29–42). New York, NY, USA: Springer New York.

Bennett, C. A. (1971). Toward empirical, practicable, comprehensive task taxonomy, *Human Factors*, 13, 229 - 235.

Berliner, D. C., Angell, D., & Shearer, J. W. (1964, August). Behaviors, measures, and instruments for performance evaluation in simulated environments. In *Symposium on the Quantification of Human Performance* (pp. 17-19).

Bernard, H. R., & Bernard, H. R. (2012). *Social research methods: Qualitative and quantitative approaches*. Sage.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74.

Black, P. & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)*, 21(1), 5-31.

Black, P., Wiliam, D. (2012) Assessment for Learning in the Classroom. In Gardner, J. N., & Gardner, J. (Eds.), *Assessment and learning*. (pp. 11 - 32). Sage.

Bloom, B. S. (Ed.), Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain, by a Committee of College and University Examiners*. London, UK: Longman.

Bloom, B.S. 1969. Some theoretical issues relating to educational evaluation. In Educational evaluation: New roles, new means. In R.W. Tyler, (Ed.), *The 63rd yearbook of the National Society for the Study of Education, part 2* (Vol. 69, pp. 26–50). Chicago, IL: University of Chicago Press.

Bloomfield, A., Deng, Y., Wampler, J., Rondot, P., Harth, D., McManus, M., Badler, N. (2003). A taxonomy and comparison of haptic actions for disassembly tasks, In J. Chen, B. Froehlich, B. Loftin, U. Neumann, & H. Takemura (Eds.), *Proceedings of Virtual Reality, Los Angeles, California* (pp. 225 - 231). Los Alamitos, CA, USA: IEEE Computer Society.

Bogdanovych, A., Rodriguez, J. A., Simoff, S. & Cohen, A. (2009, January). Virtual agents and 3D virtual worlds for preserving and simulating cultures. In *Intelligent Virtual Agents* (pp. 257-271). Springer Berlin Heidelberg.

Boud, D. (1995). Assessment and learning: contradictory or complementary. *Assessment for learning in higher education*, 35-48.

Breakey, K. M., Levin, D., Miller, I. & Hentges, K. E. (2008). The use of scenario-based-learning interactive software to create custom virtual laboratory scenarios for teaching genetics. *Genetics*, 179(3), 1151-1155.

Bricken, W. (1990). Virtual reality: Directions of growth. *Notes from the SIGGRAPH '90 panel (Technical Report R-90-1)*. Seattle: Human Interface Technology Laboratory, University of Washington.

Bronack, S. C., Cheney, A. L., Riedl, R. E. & Tashner, J. H. (2008). Designing Virtual Worlds to Facilitate Meaningful Communication: Issues, Considerations, and Lessons Learned. *Technical Communication*, 55(3), 261–269.

Bruce, B. & Bloch, N. (2012). Learning by Doing. In N. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 1821–1824). Boston, MA: Springer US.

Buche, C., Querrec, R., Loor, P. D. & Chevaillier, P. (2003, December). MASCARET: pedagogical multi-agents systems for virtual environment for training. In *Proceedings of International Conference on Cyberworlds* (pp. 423-430). IEEE.

Cacioppo, J. T., & Tassinary, L. G. (1990). Inferring psychological significance from physiological signals. *American Psychologist*, 45(1), 16-28.

Caliskan, H. (2012). Inquiry Learning, In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 1571-1573). Boston, MA: Springer US.

Cannon-Bowers, J. A., Tannenbaum, S. I., Salas, E. and Volpe, C. E. (1995). Determining team competencies: implications for training requirements and strategies. In R. Guzzo and E. Salas (Eds.), *Team Effectiveness and Decision Making in Organizations* (pp. 333-380). San Francisco, CA, USA: Jossey-Bass.

Cappella, J. N., & Pelachaud, C. (2002). Rules for responsive robots: Using human interactions to build virtual interactions. In A. L. Vangelisti, H. T. Reis, & M. A. Fitzpatrick (Eds.), *Stability and change in relationships* (pp. 325-354). Cambridge, MA, USA: Cambridge University Press.

Carlsson, C., Hagsand, O. (1993). DIVE—A platform for multi-user virtual environments, *Computers & Graphics*, 17(6), 663 - 669.

Carroll, J. M. (2000). *Making use: scenario-based design of human-computer interactions*. Cambridge, MA, USA: MIT press.

Chadwick, B., Bahr, H. M., Albrecht, S. L. (1984). *Social Science Research Methods*. NJ, USA: Prentice-Hall Inc.

Chen, C. & Rada, R. (1996). Modelling situated actions in collaborative hypertext databases. *Journal of Computer‐Mediated Communication*, 2(3).

Chodos, D., Stroulia, E., King, S. & Carbonaro, M. (2014). A framework for monitoring Instructional environments in a virtual world. *British Journal of Educational Technology*, 45(1), 24 - 35.

Cilliers, F., Schuwirth, L., Adendorff, H., Herman, N. & van der Vleuten, C. (2010). The mechanism of impact of summative assessment on medical students' learning. *Advances in Health Sciences Education*, 15(5), 695–715.

Clark, I. (2012). Formative Assessment: Assessment Is for Self-regulated Learning. *Educational Psychology Review,* 24(2), 205 - 249.

Clemen, R. T. (1989). Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting*, 5, 559–583.

Cleven, A., Gubler, P. & Hüner, K. M. (2009). Design Alternatives for the Evaluation of Design Science Research Artifacts. In *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology* (pp. 19:1–19:8). New York, USA: ACM.

Cockayne, W. R. (1998). *Two-hand, whole-hand interaction* (Doctoral dissertation, Monterey, California. Naval Postgraduate School).

Cockayne, W., Darken, R. (2004). The application of human ability requirements to virtual environment interface design and evaluation. In Dan Diaper and Neville Stanton (Eds.) *The handbook of task analysis for human-computer interaction* (pp. 401–421). Mahwah, NJ, USA: Lawrence Erlbaum Associates.

Collis, J. & Hussey, R. (2014). *Business research: A practical guide for undergraduate and postgraduate students,* Fourth Edition. Palgrave Macmillan.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*, Fourth Edition. Sage Publications.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika,* 16(3), 297-334.

D'Arcy, E. (1963). *Human Acts*. New York, NY, USA: Oxford University Press.

Davidson, D. (1967). The logical form of action sentences. In N. Rescher (Ed.), *The Logic of Decision and Action*. University of Pittsburgh Press.

Dillenbourg, P., Schneider, D. & Synteta, P. (2002). Virtual learning environments. In *3rd Hellenic Conference" Information & Communication Technologies in Education"*, (pp. 3-18). Kastaniotis Editions, Greece.

Doering, A. (2006). Adventure learning: Transformative hybrid online education. *Distance Education*, 27(2), 197-215.

Dores, A. R., Carvalho, I. P., Barbosa, F., Almeida, I., Guerreiro, S., Oliveira, B. & Caldas, A. C. (2012). Computer-Assisted Rehabilitation Program–Virtual Reality (CARP-VR): A Program for Cognitive Rehabilitation of Executive Dysfunction. In *Virtual and networked organizations, emergent technologies and tools* (pp. 90-100). Springer Berlin Heidelberg.

Dreher, C., Reiners, T. & Dreher, H. (2011). Investigating Factors Affecting the Uptake of Automated Assessment Technology. *Journal of Information Technology Education: Research*, 10(1), 161-181. Informing Science Institute.

Duncan, I., Miller, A., Jiang, S. (2012). A taxonomy of virtual worlds usage in education. *British Journal of Educational Technology*, 43(6), 949 - 964.

Earl, L. M. (2012). *Assessment as learning: Using classroom assessment to maximize student learning*. CA, USA: Corwin Press.

Edelson, D. C., Gordin, D. N. & Pea, R. D. (1999). Addressing the challenges of inquiry-based learning through technology and curriculum design. *Journal of the learning sciences*, 8(3-4), 391-450.

Efron, D. (1941). *Gesture and environment*. Oxford, UK: King'S Crown Press Gesture and environment.

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1), 49-98.

Epps, J., Lichman, S. & Wu, M. (2006). A study of hand shape use in tabletop gesture interaction. In *CHI '06 extended abstracts on human factors in computing systems* (pp. 748–753). New York, NY, USA: ACM Press.

Errington, E. P. (2003) *Developing Scenario-based Learning: practical insights for tertiary educators*. Palmerston North, New Zealand: Dunmore Press.

Eschenbrenner, B., Fui-Hoon Nah, F., Siau, F. (2008). 3-D Virtual Worlds in Education: Applications, Benefits, Issues, and Opportunities. *Journal of Database Management*, 19(4), 91-110.

Evans, J. S. B. & Over, D. E. (2013). *Rationality and reasoning*. Psychology Press.

Everitt, B. S., Landau, S., Leese, M., Stahl, D. (2011). Detecting Clusters Graphically. In *Cluster Analysis* (pp. 15–41). John Wiley & Sons, Ltd.

Experience API, Version 1.0.1. (2013, October 1). The Advanced Distributed Learning (ADL) Initiative. Retrieved from http://52.0.16.95/wp-content/uploads/2013/10/xAPI_v1.0.1-2013-10-01.pdf

Fardinpour, A., Reiners, T. (2014). The Taxonomy of Goal-oriented Actions in Virtual Training Environments, *Procedia Technology*, 13, 38 - 46.

Fardinpour, A., Reiners, T., Dreher, H. (2013). Action-based Learning Assessment Method (ALAM) in Virtual Training Environments. In M. Gosper, J. Hedberg, H. Carter (Eds.) *Electric Dreams. Proceedings Ascilite Sydney 2013* (pp. 267-276). Sydney, NSW: Macquarie University.

Farley, H. (2011). Interoperability, Learning Designs and Virtual Worlds: Issues and Strategies. In F. Lazarinis, S. Green, & E. Pearson (Eds.) *Handbook of Research on E-Learning Standards and Interoperability: Frameworks and Issues* (pp. 193-206). PA, USA: Hershey.

Fettke, P., & Loos, P. (2003). Multiperspective Evaluation of Reference Models – Towards a Framework. In Manfred A. Jeusfeld & Oscar PastorM. Jeusfeld & Pastor (Eds.), *Conceptual Modeling for Novel Application Domains* (Vol. 2814, pp. 80–91). Springer Berlin Heidelberg.

Finley, D. L., Obermayer, R. W., Bertone, C. M., Meister, D. and Muckler, F. A. (1970). *Human performance prediction in man-machine systems: A technical review*, NASA CR-1614, 1, Canoga Park, CA: The Bunker-Ramo Corporation.

Fleishman, E. A. & Reilly, M. E. (1992). *Fleishmann Job Analysis Survey*. Consulting Psychologists Press.

Fleishman, E. A. (1975). Toward a taxonomy of human performance. *American Psychologist*, 30(12), 1127 - 1149.

Fleishman, E. A. and Mumford, M. D. (1991), Evaluating Classifications of Job Behavior: A Construct Validation of the Ability Requirement Scales. *Personnel Psychology*, 44(3), 523–575.

Fleishman, E. A., Quaintance, M. K., Broedling, L. A. (1984). *Taxonomies of human performance: The description of human tasks* (pp. xvi 514), San Diego, CA, US: Academic Press.

Foltz, P. W., Hidalgo, P. & Van Moere, A. (2014a). Improving Student Writing through Automated Formative Assessment: Practices and Results. *International Association for Educational Assessment (IAEA) 2014 Conference*, 1-18.

Foltz, P. W., Rosenstein, M., Dronen, N. & Dooley, S. (2014b). Automated feedback in a large-scale implementation of a formative writing system: Implications for improving student writing. *Presented at the American Educational Research Association Annual Meeting*.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383), 553-569.

Francis, J., Johnston, M., Robertson, C., Glidewell, L., Entwistle, V., Eccles, M. P. & Grimshaw, J. M. (2010). What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychology & Health*, 25(10), 1229-1245.

Frank, U. (2007). Evaluation of reference models. *Reference modelling for business systems analysis*, 118-140.

Freedman, N., & Hoffman, S. P. (1967). Kinetic behavior in altered clinical states: Approach to objective analysis of motor behavior during clinical interviews. *Perceptual and motor skills*, 24(2), 527-539.

Freeman, D., Benko, H., Morris, M. R. & Wigdor, D. (2009, November). ShadowGuides: visualizations for in-situ learning of multi-touch and whole-hand gestures. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces* (pp. 165-172). ACM.

Frey, B. B., & Schmitt, V. L. (2007). Coming to terms with classroom assessment. *Journal of Advanced Academics*, 18(3), 402-423.

Frey, B. B., Schmitt, V. L. & Allen, J. P. (2012). Defining authentic classroom assessment. *Practical Assessment, Research & Evaluation*, 17(2), 1-18.

Fuchs, P., Guitton, P. (2011). Introduction to virtual reality. In Philippe Fuchs, Guillaume Moreau, Pascal Guitton (Ed.), *Virtual Reality: Concepts and Technologies* (pp. 3 - 10). Boca Raton, FL, USA: CRC Press.

Fujita, Y., Hollnagel, E. (2004). Failures without errors: quantification of context in HRA, *Reliability Engineering & System Safety*, 83 (2), 145-151.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2), 155-170.

George, D., and Mallery, P. (2013). *IBM SPSS Statistics 21 Step by Step: A Simple Guide and Reference (13 edition)*. Pearson.

Gerbaud, S. & Arnaldi, B. (2008, October). Scenario sharing in a collaborative virtual environment for training. In *Proceedings of the 2008 ACM symposium on Virtual Reality Software and Technology* (pp. 109-112). ACM.

Gerbaud, S., Mollet, N. & Arnaldi, B. (2007). Virtual Environments for Training: From Individual Learning to Collaboration with Humanoids. In K. Hui, Z. Pan, R. Chung, C. L. Wang, X. Jin, S. Göbel, & E.-L. Li (Eds.), *Technologies for E-Learning and Digital Entertainment* (Vol. 4469, pp. 116–127). Springer Berlin Heidelberg.

Gesture [Def. 1]. (2016). Oxford Advanced Learner's Dictionary. Retrieved from http://www.oxfordlearnersdictionaries.com/definition/english/gesture

Gijbels, D., Dochy, F., Van den Bossche, P. & Segers, M. (2005). Effects of problem-based learning: A meta-analysis from the angle of assessment. *Review of Educational Research*, 75(1), 27-61.

Gikandi, J.W., Morrow, D., Davis, N.E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education*, 57(4), 2333-2351.

Given, L. M. (Ed.). (2008). *The Sage encyclopedia of qualitative research methods*. Sage Publications.

Goldman, A. I. (1970) *A Theory of Human Action* (pp. 230), Upper Saddle River, NJ, USA: Prentice-Hall.

Goulding, J., Nadim, W., Petridis, P. & Alshawi, M. (2012). Construction industry offsite production: A virtual reality interactive training environment prototype. *Advanced Engineering Informatics*, 26(1), 103-116.

Graf, S., Kinshuk & Liu, T. (2009). Supporting teachers in identifying students' learning styles in learning management systems: An automatic student modelling approach. *Journal of Educational Technology & Society*, 12(4), 3-14.

Gregor, S., & Jones, D. (2007). The anatomy of a design theory. *Journal of the Association for Information Systems*, 8(5), 312-335.

Gregory, S. & Masters, Y. (2012). Real thinking with virtual hats: A roleplaying activity for pre-service teachers in Second Life. In M. J. W. Lee, B. Dalgarno & H. Farley (Eds.), *Virtual worlds in tertiary education: An Australasian* *perspective*. Australasian Journal of Educational Technology, 28(Special issue, 3), 420-440.

Gregory, S., Gregory, B., Reiners, T., Hillier, M., Lee, M. J.W., Jacka, L., Larson, I. (2013). Virtual worlds in Australian and New Zealand higher education: Remembering the past, understanding the present and imagining the future. In H. Carter, M. Gosper and J. Hedberg (Eds.), *Electric Dreams. Proceedings ascilite 2013 Sydney*. (pp. 312- 324).

Grobe, S. J. & Hughes, L. C. (1993). The conceptual validity of a taxonomy of nursing interventions. *Journal of Advanced Nursing*, 18(12), 1942-1961.

Guest, G., Bunce, A. & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods*, 18(1), 59-82.

Gulikers, J. T., Bastiaens, T. J. & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development*, 52(3), 67-86.

Gulikers, J. T., Bastiaens, T. J. & Martens, R. L. (2005). The surplus value of an authentic learning environment. *Computers in Human Behavior*, 21(3), 509-521.

Halkidi, M., Batistakis, Y., Vazirgiannis, M. (2002). Cluster validity methods: part I, SIGMOD Rec, 31 (2), 40-45.

Hannah, S. (2008). *Sorting out card sorting: Comparing methods for information architects, usability specialists, and other practitioners*. Retrieved from http://hdl.handle.net/1794/7818

Hargreaves, E., Gipps, C. & Pickering, A. (2014). *Assessment for learning* (p. 313). UK: Angela McLachlan, University of Manchester.

Hart, D. (1994). *Authentic assessment: A handbook for education*. Menlo Park, CA, USA: Addison-Wesley Publishing Company.

Hattie, J., Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1), 81 – 112.

Heinrichs, W. L., Youngblood, P., Harter, P. M. & Dev, P. (2008). Simulation for team training and assessment: case studies of online training with virtual worlds. *World journal of surgery*, 32(2), 161-170.

Herrington, A. & Herrington, J. (2008). What is an Authentic Learning Environment? In L. Tomei (Ed.), *Online and Distance Learning: Concepts, Methodologies, Tools, and Applications* (pp. 68-77). Hershey, PA

Herrington, J. & Herrington, A. (1998). Authentic assessment and multimedia: How university students respond to a model of authentic assessment. *Higher Education Research & Development*, 17(3), 305-322.

Herron, M. D. (1971). The Nature of Scientific Enquiry. *The School Review*, 171-212.

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian journal of information systems*, 19(2), Article 4.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 28(1), 75-105.

Hogan, J., Broach, D., & Salas, E. (1990). Development of a task information taxonomy for human performance systems. *Military Psychology*, 2(1), 1-19.

Holding, D. H. (1989). *Human Skills*. New York, NY, USA: John Wiley & Sons.

Holstein, W. K. (n.d.). Human-factors engineering. In *Encyclopaedia Britannica*.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, 15(3), 495-514.

Hu, X., & Xia, H. (2010). Automated Assessment System for Subjective Questions Based on LSI. In *Third International Symposium on Intelligent Information Technology and Security Informatics (IITSI)*, 250-254. IEEE.

Hung, W., Jonassen, D. H., & Liu, R. (2008). Problem-based learning. In M. Spector, D. Merrill, J. van Merrienboer, & M. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 485–506). New York, NY, USA: Erlbaum.

Hurst, S. (2011). Towards high-resolution ethnography for evaluation of team-oriented virtual reality training for medicine. Presentation given at *Medicine Meets Virtual Reality 18*, Newport Beach, CA, USA, February 9–12, 2011.

Hwang, G.J. & Chang, H.F. (2011). A formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students. *Computers & Education*, 56(4), 1023 – 1031.

Iivari, J. (2003). The IS core-VII: Towards information systems as a science of meta-artifacts. *Communications of the Association for Information Systems*, 12(1), 37.

Iivari, J. (2007). A paradigmatic analysis of information systems as a design science. *Scandinavian journal of information systems*, 19(2), Article 5.

Iverson, K. & Colky, D. (2004). Scenario-based E-learning design. *Performance Improvement*, 43(1), 16-22.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New phytologist*, 11(2), 37-50.

Jarmon, L., Traphagan, T., Mayrath, M., & Trivedi, A. (2009). Virtual world teaching, experiential learning, and assessment: An interdisciplinary communication course in Second Life. *Computers & Education*, 53(1), 169-182.

Jonassen, D. H., & Hung, W. (2012). Problem-Based Learning. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 2687–2690). Boston, MA: Springer US.

Karam, M. and Schraefel, m. c. (2005). *A taxonomy of gestures in human computer interactions*. UK: University of Southhampton. http://eprints.soton.ac.uk/id/eprint/261149

Kervin, J. B. (1992). *Methods for business research*. Harper Collins.

Kim, J.W. & Jung, W. D. (2003). A taxonomy of performance influencing factors for human reliability analysis of emergency tasks. *Journal of Loss Prevention in the Process Industries*, 16 (6), 479–495.

Klastrup, L. (2003). A Poetics of Virtual Worlds. In *Proceedings of the Fifth International Digital Arts and Culture Conference*, Melbourne, Australia: RMIT School of Applied Communication.

Kneebone, R. L., Kidd, J., Nestel, D., Barnet, A., Lo, B., King, R. & Brown, R. (2005). Blurring the boundaries: scenario-based simulation in a clinical setting. *Medical Education*, 39(6), 580-587.

Knight, P. (Ed.). (2012). *Assessment for learning in higher education*. OX, UK: Routledge.

Kolb, A. Y., & Kolb, D. A. (2010a). *Experiential learning theory bibliography: 1971–2005*. Cleveland: Experience Based Learning Systems. www.learningfromexperience.com.

Kolb, A. Y., & Kolb, D. A. (2010b). *Experiential learning theory bibliography: Recent research 2006–2010*. Cleveland: Experience Based Learning Systems. www.learningfromexperience.com.

Kolb, A., & Kolb, D. (2012). Experiential Learning Theory. In N. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 1215–1219). Boston, MA: Springer US.

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Englewood Cliffs: Prentice-Hall.

Kopainsky, B., Pirnay-Dummer, P. & Alessi, S. M. (2012). Automated assessment of learners' understanding in complex dynamic systems. *System Dynamics Review*, 28(2), 131-156.

Krueger, M. (1991). *Artificial Reality II*. Reading, MA, USA: Addison-Wesley Professional.

Kuechler, B. & Vaishnavi, V. (2008). On theory development in design science research: anatomy of a research project. *European Journal of Information Systems*, 17(5), 489-504.

Kuechler, W. & Vaishnavi, V. (2012). A framework for theory development in design science research: multiple perspectives. *Journal of the Association for Information systems*, 13(6), 395-423.

Kumar, R. (2014). *Research Methodology A Step-by-Step Guide for Beginners (FOURTH EDITION)*. SAGE Publications Ltd.

Kuutti, K. (1995) Activity Theory as a potential framework for human-computer interaction research, In B. Nardi (ed.) *Context and Consciousness: Activity Theory and Human Computer Interaction* (pp. 17-44). Cambridge, MA, USA: MIT Press.

Kuzel, AJ. (1999). Sampling in qualitative inquiry. In B.F. Crabtrree and W.L. Miller (Eds.) *Doing Qualitative Research* (second edition)(pp. 33-45). Thousand Oaks, CA: Sage Publlications.

Lasky, B., & Tempone, I. (2004). Practising what we teach: vocational teachers learn to research through applying action learning techniques. *Journal of further and higher education,* 28(1), 79-94.

Lester, J., Choudhury, T., Kern, N., Borriello, G., & Hannaford, B. (2005, July). A hybrid discriminative/generative approach for modelling human activities. In *Nineteenth International Joint Conference on Artificial Intelligence*, Edinburgh, Scotland (Vol. 5, pp. 766-772).

Levy, P. S., & Lemeshow, S. (2013). *Sampling of populations: methods and applications*, 4th Edition. John Wiley & Sons.

Lincoln, Y. S., Lynham, S. A. & Guba, E. G. (2011). Paradigmatic controversies, contradictions, and emerging confluences revisited. *The Sage handbook of qualitative research*, 4, 97-128.

Logan, A., Stuart, R. (1987). Action Based Learning: Are Activity and Experience the Same? *Industrial and Commercial Training*, 19(2), 16 - 20.

Loh, C. S., & Sheng, Y. (2014). Maximum Similarity Index (MSI): A metric to differentiate the performance of novices vs. multiple-experts in serious games. *Computers in Human Behavior*, 39, 322-330.

Mackenzie, N. & Knipe, S. (2006). Research dilemmas: Paradigms, methods and methodology. *Issues in educational research*, 16(2), 193-205.

Mackieh, A., Cilingir, C. (1998). Effects of performance shaping factors on human error, *International Journal of Industrial Ergonomics*, 22 (4–5), 285-292.

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition*, 1(1), 18-26.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press. pp. 423.

McTighe, J. and O'Connor, K. (2005). Seven Practices for Effective Learning. *Educational Leadership*, 63(3), 10 - 17.

Méndez, G., de Antonio, A. & Herrero, P. (2001). Prvir: An integration between an intelligent tutoring system and a virtual environment. In *SCI2001*, 8, 175-180.

Méndez, G., Herrero, P., & de Antonio, A. (2004, April). Intelligent Virtual Environments for Training in Nuclear Power Plants. In *6th International Conference on Enterprise Information Systems* (pp. 204-209). Porto, Portugal: Universidade Portucalense

Méndez, G., Rickel, J. & de Antonio, A. (2003). Steve Meets Jack: The Integration of an Intelligent Tutor and a Virtual Environment with Planning Capabilities. In

T. Rist, R. Aylett, D. Ballin, & J. Rickel (Eds.), *Intelligent Virtual Agents* (Vol. 2792, pp. 325–332). Springer Berlin Heidelberg.

Mertens, D. M. (2015). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods*, Fourth Edition. Sage Publications.

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., Wood, C. (2013). The Behavior Change Technique Taxonomy (v1) of 93 Hierarchically Clustered Techniques: Building an International Consensus for the Reporting of Behavior Change Interventions. *Annals of Behavioural Medicine*, 46(1), 81–95.

Mills, C. W., (1940). Situated Actions and Vocabularies of Motive. *American Sociological Review*, 5(6), 904–913.

Minton, E. A., Khale, L. R. (2014). *Belief Systems, Religion, and Behavioural Economics*. New York: Business Expert Press LLC.

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., Johnson, L. (1999). A cognitive task analysis with implications for designing simulation-based performance assessment, *Computers in Human Behavior*, 15(3–4), 335 - 374.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., … de Vet, H. C. W. (210). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, 63(7), 737–745.

Moos, D. C. & Honkomp, B. (2011). Adventure learning: Motivating students in a Minnesota middle school. *Journal of Research on Technology in Education*, 43(3), 231-252.

Moraes, R. M., Machado, L. S., & Souza, L. C. (2012). Skills Assessment of Users in Medical Training Based on Virtual Reality Using Bayesian Networks. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* (pp. 805-812). Springer Berlin Heidelberg.

Morie, J. F. (2007). Performing in (virtual) spaces: Embodiment and being in virtual environments. *International Journal of Performance Arts and Digital Media*, 3(2-3), 123-138.

Moskaliuk, J., Bertram, J., Ulrike Cress, U. (2013). Training in virtual environments: putting theory into practice. *Ergonomics*, 56(2), 195-204.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3), 691-692.

Naidu, S. (2010). Using scenario-based learning to promote situated learning and develop professional knowledge. In E. Errington (Ed.), *Preparing graduates for the professions using scenario-based learning*, (pp. 39-49). Mt Gravatt, QLD, Australia: Post Pressed.

Naidu, S., & Bedgood, D. R. (2012). Action-based Learning. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 75–77). Boston, MA: Springer US.

NCroarkin, C., & Tobias, P. (2015). *NIST/SEMATECH e-handbook of statistical methods. NIST/SEMATECH*, July. Available online: http://www.itl.nist.gov/div898/handbook.

Nelson, B. C., Kim, Y., Foshee, C., Slack, K. (2014). Visual signalling in virtual world-based assessments: The SAVE Science project. *Information Sciences*, 264, 32-40.

Novikov, A. M., & Novikov, D. A. (2013). *Research methodology: From philosophy of science to research design (Vol. 2).* Boca Raton, FL, USA: CRC Press.

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: what can be done? *Assessment & Evaluation in Higher Education*, 33(3), 301-314.

Nunamaker Jr, J. F., Chen, M., & Purdin, T. D. (1990). Systems development in information systems research. *Journal of management information systems*, 7(3), 89-106.

Offermann, P., Levina, O., Schönherr, M., & Bub, U. (2009, May). Outline of a design science research process. In *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology*, Article 7. ACM.

Oller, R. G. (1968). Human factors data thesaurus: an application to task data (AD-670 -578), *Technical Paper for the Air Force* (p. 71). Dayton, Ohio, USA: Systems Development Corp.

Ong, J. (2007). Automated performance assessment and feedback for free-play simulation-based training. *Performance Improvement*, 46(10), 24-31.

Palmer, J., Williams, R., & Dreher, H. (2002). Automated essay grading system applied to a first year university subject–How can we do it better. In *Proceedings of Informing Science 2002 Conference*, Cork, Ireland, June (pp. 19-21).

Panait, L., Hogle, N. J., Fowler, D. L., Bell, R. L., Roberts, K. E., Duffy, A. J. (2011). Completion of a Novel, Virtual-Reality-Based, Advanced Laparoscopic Curriculum Improves Advanced Laparoscopic Skills in Senior Residents. *Journal of Surgical Education*, 68(2), 121 - 125.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. In *Proceedings of the Royal Society of London*, 58, 240-242.

Peffers, K., Rothenberger, M., Tuunanen, T., & Vaezi, R. (2012). Design science research evaluation. In *Design science research in information systems. Advances in theory and practice* (pp. 398-410). Springer Berlin Heidelberg.

Peffers, K., Tuunanen, T., Gengler, C. E., Rossi, M., Hui, W., Virtanen, V., & Bragge, J. (2006, February). The design science research process: a model for producing and presenting information systems research. In *Proceedings of the* *first international conference on design science research in information systems and technology (DESRIST 2006)* (pp. 83-106).

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, 24(3), 45-77.

Perrenot, C., Perez, M., Tran, N., Jehl, J. P., Felblinger, J., Bresler, L., & Hubert, J. (2012). The virtual reality simulator DV-Trainer® is a valid assessment tool for robotic surgical skills. *Surgical endoscopy*, 26(9), 2587-2593.

Pfeiffer, D., & Niehaves, B. (2005). Evaluation of conceptual models-a structuralist approach. *ECIS 2005 Proceedings*, 43.

Phelan, C., & Wren, J. (2006). *Exploring reliability in academic assessment*. UNI Office of Academic Assessment.

Pirsiavash, H., Ramanan, D., (2012). Detecting activities of daily living in first-person camera views, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, USA. 2847 - 2854.

Piumsomboon, T., Clark, A., Billinghurst, M., & Cockburn, A. (2013). User-defined gestures for augmented reality. In *Human-Computer Interaction–INTERACT 2013* (pp. 282-299). Springer Berlin Heidelberg.

Popham, W.J. 2008. *Transformative assessment*. Alexandria, VA, USA: ASCD.

Pries-Heje, J., & Baskerville, R. (2008). The design theory nexus. *MIS Quarterly*, 731-755.

Pries-Heje, J., Baskerville, R., & Venable, J. (2008). Strategies for design science research evaluation. *ECIS 2008 proceedings*, 1-12.

Project Management Institute (2013). *A Guide to the Project Management Body of Knowledge (PMBOK® Guide)*-Fifth Edition. Newtown Square, PA, USA: Project Management Institute

Purao, S. (2002). *Design Research in the Technology of Information Systems: Truth or Dare*. GSU Department of CIS Working Paper. Atlanta.

Reed, L. E. (1967). *Advances in the use of computers for handling human factors task data (No. 670291)*. SAE Technical Paper.

Reeves, T., Herrington, J. & Oliver, R. (2002) Authentic activities and online learning, in Quality Conversations, In *Proceedings of the 25th HERDSA Annual Conference, Perth, Western Australia* (pp. 562-567). Milperra, NSW, Australia: Higher Education Research and Development Society of Australasia, Inc.

Reiners, T., Teräs, H., Chang, V., Wood, L. C., Gregory, S., Gibson, D., Petter, N. & Teräs, M. (2014). Authentic, immersive, and emotional experience in virtual learning environments: The fear of dying as an important learning experience in a simulation. In *Transformative, innovative and engaging. Proceedings of the 23rd Annual Teaching Learning Forum*, 30-31 January 2014. Perth: The University of Western Australia.

http://ctl.curtin.edu.au/professional_development/conferences/tlf/tlf2014/refereed/reiners.html

Rickel, J. and Johnson, W. L. (1998). STEVE (video session): a pedagogical agent for virtual reality. In Katia P. Sycara and Michael Wooldridge (Eds.) *Proceedings of the second international conference on Autonomous agents* (pp. 332 – 333). New York, NY, USA: ACM.

Rickel, J., & Johnson, W. L. (1999). Virtual humans for team training in virtual reality. In *Proceedings of the ninth international conference on artificial intelligence in education*. (pp. 578-585). Netherlands: IOS Press

Roberts, T. S. (2006). The Use of Multiple Choice Tests for Formative and Summative Assessment. *In Proceedings of the 8th Australasian Conference on Computing Education* (Vol. 52, pp. 175–180). Darlinghurst, Australia: Australian Computer Society, Inc.

Robertson, T. (1997). Cooperative work and lived cognition: a taxonomy of embodied actions. In *Proceedings of European conferences on Computer-Supported Cooperative Work*, 97, 205 - 220.

Robertson, T. (2000). Building bridges: negotiating the gap between work practice and technology design. *International Journal of Human-Computer Studies*, 53(1), 121 - 146.

Rogers, C. (1951). *Client-Centred Therapy*, London, UK: Constable.

Roland, J., Figueira, J. R., & De Smet, Y. (2016). Finding compromise solutions in project portfolio selection with multiple experts by inverse optimization. *Computers & Operations Research*, 66, 12-19.

Rose, F. D., Attree, E. A., Brooks, B. M., Parslow, D. M., Penn, P. R. (2000). Training in Virtual Environments: Transfer to Real World Tasks and Equivalence to Real Task Training. *Ergonomics*, 43(4), 494–511.

Rossi, M. and Sein, M. (2003). Design Research Workshop: A Proactive Research Approach. In *Presentation delivered at IRIS* 26, August 9 – 12, 2003.

Rosson, M. B., & Carroll, J. M. (2009). Scenario based design. In A. Sears & J.A. Jacko (Eds), *Human‐computer interaction: Development Process* (pp. 145-162). Boca Raton, FL: CRC Press.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional science*, 18(2), 119-144.

Salmela, L., & Tarhio, J. (2004). ACE: Automated compiler exercises. In *Proceedings of the 4th Finnish/Baltic sea conference on computer science education*. 131-135.

Salmon, P., Stanton, N., Baber, C., Walker, G., Green, D. (2004). *Human factors design & evaluation methods review (HFIDTC/1.3.3/1-1)*, Human Factors Integration Defence Technology Centre Report.

Salmon, P., Stanton, N., Walker, G., Jenkins, D. P. (2008). *Rapid Assessment of Tasks & Context (RATaC): Methodological Development*

(HFIDTC/2/WP5.2.3/1), BAE Systems, Issued by Aerosystems International Ltd on behalf of the HFI DTC consortium.

Savery, J. R. (2006). Overview of Problem-based Learning: Definitions and Distinctions. *Interdisciplinary Journal of Problem-Based Learning,* 1(1), 9-20.

Schramma, E., & Srinivasan, V. (2015). *WritingAssistant™ Comprehensive automated feedback*. EnglishHelper, Inc.

Schwab, J. J. (1960). Inquiry, the science teacher, and the educator. *The School Review*, 176-195.

Scriven, M. (1967). The Methodology of Evaluation. In R.W. Tyler, R.M. Gagne, M. Scriven (Eds.), *AERA Monograph Series, Perspectives of Curriculum Evaluation,* Chicago, IL, USA: Rand McNally.

Segers, M., & Dochy, F. (2001). New assessment forms in problem-based learning: the value-added of the students' perspective. *Studies in higher education*, 26(3), 327-343.

Shaffer, D. W., & Resnick, M. (1999). "Thick" Authenticity: New Media and Authentic Learning. *Journal of interactive learning research,* 10(2), 195-215.

Shannon, D. M., & Bradshaw, C. C. (2002). A comparison of response rate, response time, and costs of mail and electronic surveys. *The Journal of Experimental Education*, 70(2), 179-192.

Shannon, D. M., Johnson, T. E., Searcy, S., & Lott, A. (2002). Using electronic surveys: Advice from survey professionals. *Practical Assessment, Research & Evaluation*, 8(1), 1-2.

Shen, R., Tang, Y., Zhang, T. Z. (2001). The intelligent assessment system in Web-based distance learning education. *31st Annual Conference of Frontiers in Education*, 1, T1F 7 - 11.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. *International encyclopedia of education*, 4, 20-26.

Shih, H. S., Huang, L. C., & Shyur, H. J. (2005). Recruitment and selection processes through an effective GDSS. *Computers & Mathematics with Applications*, 50(10), 1543-1558.

Shute, V. J. (2007). Focus on Formative Feedback. *ETS Research Report Series*, 2007(1), i–47.

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). Mahwah, NJ: Routledge, Taylor and Francis.

Shute, V., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. Cambridge, MA: The MIT Press

Shwayder, D. S. (1965). *The Stratification of Behaviour*. New York, NY, USA: Humanities Press.

Siau, K., & Rossi, M. (2011). Evaluation techniques for systems analysis and design modelling methods – a review and comparative analysis. *Information Systems Journal*, 21(3), 249–268.

Simpson, G. G. (1961). *Principles of animal taxonomy* (No. 20). Columbia University Press.

Sluijsmans, D. M., Moerkerke, G., Van Merrienboer, J. J., & Dochy, F. J. (2001). Peer assessment in problem based learning. *Studies in educational evaluation*, 27(2), 153-173.

Smee, A., & Brennan, M. (2000, November). Electronic surveys: A comparison of e-mail, web and mail. In *Proceedings of ANZMAC*.

Sonnenberg, C., & vom Brocke, J. (2012). Evaluation patterns for design science research artefacts. In *Practical Aspects of Design Science* (pp. 71-83). Springer Berlin Heidelberg.

Spencer, D. (2009). *Card sorting: Designing usable categories*. Rosenfeld Media.

Steuer, J. (1992). Defining Virtual Reality: Dimensions Determining Telepresence. *Journal of Communication*, 42, 73–93.

Stone, R. J. (2004). Rapid assessment of tasks and context (RATaC) for technology-based training, In *Proceedings of I/ITSEC 2004* (pp. 6–9), Orlando, Florida.

Suchman, L. (2007). *Human-machine reconfigurations: Plans and situated actions*. Cambridge University Press.

Swaffield, S. (2011). Getting to the heart of authentic assessment for learning. *Assessment in Education: Principles, Policy & Practice*, 18(4), 433-449.

Swezey, R. W., Owens, J. M., Bergondy, M. L., Salas, E. (1998). Task and training requirements analysis methodology (TTRAM): an analytic methodology for identifying potential training uses of simulator networks in teamwork-intensive task environments, *Ergonomics*, 41(11), 1678-1697.

Taras, M. (2005). Assessment: summative and formative – some theoretical reflections. *British Journal of Educational Studies*, 53(4), 466-478.

Teach [Def. 1]. (2016). Oxford Advanced Learner's Dictionary. Retrieved from http://www.oxfordlearnersdictionaries.com/definition/english/teach

Teach [Def. 2]. (2016). Oxford Advanced Learner's Dictionary. Retrieved from http://www.oxfordlearnersdictionaries.com/definition/english/teach

Teaching and Learning Services (2014). *Guidelines for assessment of experiential learning*. Montreal, Canada: Teaching and Learning Services, McGill University.

Tichon, J. G. (2007). Using presence to improve a virtual training environment. *Cyberpsychology & Behavior*, 10(6), 781-788.

Toranj, S., & Ansari, D. N. (2012). Automated Versus Human Essay Scoring: A Comparative Study. *Theory and Practice in Language Studies*, 2(4), 719-725.

Torrance, H. (1995). *Evaluating authentic assessment*. Buckingham, UK: Open University Press.

Train [Def. 1]. (2016). Oxford Advanced Learner's Dictionary. Retrieved from http://www.oxfordlearnersdictionaries.com/definition/english/train_2

Train [Def. 2]. (2016). Oxford Advanced Learner's Dictionary. Retrieved from http://www.oxfordlearnersdictionaries.com/definition/english/train_2

Trotter, E. (2006). Student perceptions of continuous summative assessment. *Assessment & Evaluation in Higher Education*, 31(5), 505-521.

Tuomela, R. (1977). *Human action and its explanation: A study on the philosophical foundations of psychology* (Vol. 116, p. 426), Springer Science & Business Media.

Turaga, P., Chellappa, R., Subrahmanian, V. S., Udrea, O. (2008). Machine Recognition of Human Activities: A Survey, *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11), 1473-1488.

Turner, K., Roberts, L., Heal, C., Wright, L. (2013). Oral presentation as a form of summative assessment in a master’s level PGCE module: the student perspective. *Assessment & Evaluation in Higher Education*, 38(6), 662-673.

Twining, P. (2009). Exploring the educational potential of virtual worlds-Some reflections from the SPP. *British Journal of Educational Technology*, 40(3), 496-514.

Urakami, J. (2012). Developing and Testing a Human-Based Gesture Vocabulary for Tabletop Systems, *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 54 (4), 636 - 653.

Vaishnavi, V. & Kuechler, W. (2004). *Design research in information systems*. January 20, 2004; last updated: November 15, 2015. URL: http://www.desrist.org/design-research-in-information-systems/

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, 2(1), 319-330.

Van Sickle, K., Buck, L., Willis, R., Mangram, A., Truitt, M., Shabahang, M., Thomas, S., Trombetta, L., Dunkin, B., Scott, D. (2011). A multicenter, simulation-based skills training collaborative using shared GI mentor II systems: results from the Texas association of surgical skills laboratories (TASSL) flexible endoscopy curriculum. *Surgical Endoscopy*, 25(9), 2980–2986.

Veletsianos, G., & Kleanthous, I. (2009). A review of adventure learning. *The International Review of Research in Open and Distributed Learning*, 10(6), 84-105.

Venable, J. (2006). The role of theory and theorising in design science research. In *Proceedings of the 1st International Conference on Design Science in Information Systems and Technology* (pp. 1-18). Claremont, CA, USA: Claremont Graduate University.

Venable, J., Pries-Heje, J., & Baskerville, R. (2012). A Comprehensive Framework for Evaluation in Design Science Research. In K. Peffers, M. Rothenberger, & B. Kuechler (Eds.), *Design Science Research in Information Systems. Advances in Theory and Practice* (Vol. 7286, pp. 423–438). Springer Berlin Heidelberg.

Verhulsdonck, G., & Morie, J. (2009). Virtual Chironomia: Developing Standards for Non-verbal Communication in Virtual Worlds. *Journal of Virtual Worlds Research*, 2(3), 1 – 10.

Vujošević-Janičić, M., Nikolić, M., Tošić, D., Kuncak, V. (2013). Software verification and graph similarity for automated evaluation of students' assignments, *Information and Software Technology*, 55(6), 1004 - 1016.

Walker, A., & Shelton, B. E. (2008). Problem-based educational games: Connections, prescriptions, and assessment. *Journal of Interactive Learning Research*, 19 (4), 663–684.

Weber, S. (2010). Design Science Research: Paradigm or Approach? In *Americas Conference on Information Systems* (*AMCIS)*, Paper 214.

Weinberg, R. S., & Gould, D. (2014). *Foundations of Sport and Exercise Psychology*, 6E. USA: Human Kinetics.

Whitlock, B., & Nanavati, J. (2013). A systematic approach to performative and authentic assessment. *Reference Services Review*, 41(1), 32 - 48.

Wiggins, G. (1990). *The Case for Authentic Assessment*. Washington, DC, USA: ERIC Clearinghouse on Tests Measurement and Evaluation, American Institutes for Research.

Wild, F., Scott, P., Lefrere, P., Karjalainen, J., Helin, K., Naeve, A., & Isaksson, E. (2014). Towards data exchange formats for learning experiences in manufacturing workplaces. In *CEUR Workshop Proceedings* (Vol. 1238, pp. 23-33).

Wilhelm, J. D., & Wilhelm, P. J. (2010). Inquiring minds learn to read, write, and think: Reaching all learners through inquiry. *Middle school journal*, 41(5), 39-46.

Williams, R., & Dreher, H. (2004). Automatically grading essays with Markit©. *Journal of Issues in Informing Science and Information Technology*, 1(2004), 693-700.

Willis, M. P. (1961). *Deriving training device implications from learning theory principles*. US Naval Training Device Center.

Wilson, B. G. (1996). *Constructivist learning environments: Case studies in instructional design*. Englewood Cliffs, NJ, USA: Educational Technology.

Winkler, R. L., & Clemen, R. T. (2004). Multiple experts vs. multiple methods: Combining correlation assessments. *Decision Analysis*, 1(3), 167-176.

Wobbrock, J. O., Morris, M. R., & Wilson, A. D. (2009). User-defined gestures for surface computing. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems* (pp. 1083–1092). New York, NY, USA: ACM Press.

Xia, P., Lopes, A. M., Restivo, M. T., & Yao, Y. (2012). A new type haptics-based virtual environment system for assembly training of complex products. *The International Journal of Advanced Manufacturing Technology*, 58(1-4), 379-396.

Xu, Z., Lu, X. Z., Guan, H., Chen, C., & Ren, A. Z. (2014). A virtual reality based fire training simulator with smoke hazard assessment capacity. *Advances in engineering software*, 68, 1-8.

Zhang, J.R.; Kuangye Guo; Herwana, C.; Kender, J.R. (2010). Annotation and taxonomy of gestures in lecture videos, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, San Francisco, CA, USA. 1-8.

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.

Appendix 1: Survey questionnaire

Area of expertise:
Have you ever used 3D virtual environments?
Yes (1)
No (2)
To what extent have you used virtual environments?
None (1)
Hobby (2)
Personal Curiosity (3)
Occasional user (4)
Frequent User (5)
Q1 How important is it to have a clear goal to achieve in any 'Action-based Learning' scenario?*
Very unimportant (1)
unimportant (2)
somewhat important (3)
important (4)
Very important (5)
Q2 How important is it to break down the solution (in any 'Action-based Learning' scenario) into different milestones? *
Very unimportant (1)
unimportant (2)
somewhat important (3)
important (4)
Very important (5)
Q3 How important is it to break down the milestones (in any 'Action-based Learning' scenario) into basic actions? *
Very unimportant (1)
unimportant (2)
somewhat important (3)
important (4)
Very important (5)
Q4 How important is it to recognise the locomotion of the avatars as an action, to assess the trainees’ performance in any 'Action-based Learning' scenario? *
Very unimportant (1)
unimportant (2)
somewhat important (3)
important (4)
Very important (5)
Q5 How important is it to recognise different types of locomotive actions in any 'Action-based Learning' scenario? *
Very unimportant (1)
unimportant (2)
somewhat important (3)
important (4)
Very important (5)

Q6 How important is it to recognise the trainees’ response as an actions to assess the trainees’ performance in any 'Action-based Learning' scenario? *
Very unimportant (1)
unimportant (2)
somewhat important (3)
important (4)
Very important (5)
Q7 How important is it to recognise the difference of trainees’ response from actions with different purposes? *(Pushing the button with the green light vs. pushing a button)
Very unimportant (1)
unimportant (2)
somewhat important (3)
important (4)
Very important (5)
Q8 How important is it to recognise those trainees' actions which are reflecting their decisions, to assess the trainees’ performance in any 'Action-based Learning' scenario? *
Very unimportant (1)
unimportant (2)
somewhat important (3)
important (4)
Very important (5)
Q9 How important is it to recognise the operative actions of trainees to assess their performance in any 'Action-based Learning' scenario? *
Very unimportant (1)
unimportant (2)
somewhat important (3)
important (4)
Very important (5)
Q10 How important is it to differentiate the actions which are not changing the objects (like breaking them) or the environment with the actions that are changing the structure or the nature of the objects? *
Very unimportant (1)
unimportant (2)
somewhat important (3)
important (4)
Very important (5)
Q11 How important is it to recognise the gestures of avatars as a communication method, to assess the trainees’ performance in any 'Action-based Learning' scenario? *
Very unimportant (1)
unimportant (2)
somewhat important (3)
important (4)
Very important (5)
Q12 How important is it to assess the trainees’ performance based on multiple experts’ solutions instead of one expert? *
Very unimportant (1)
unimportant (2)

somewhat important (3)
important (4)
Very important (5)
Q13 Most assessments are performed by one assessor, but the Action-based Learning Assessment Methodology uses a panel of assessors instead. What number of experts do you think is more suitable as a panel of reference solution providers? * Action-based Learning Assessment Methodology (ALAM) is a formative assessment method in virtual training environments, assessing learners' goal-oriented actions and action-sequences and providing them with formative feedback.
1 (1)
1-3 (2)
3-5 (3)
5 and more (4)
Q14 How important is to have a clear classification of actions in Action-based Learning Assessment? *
Very unimportant (1)
unimportant (2)
somewhat important (3)
important (4)
Very important (5)
Q15 To what extent are you familiar with the use of taxonomies? * You can find the definition of taxonomy in this link: http://en.wikipedia.org/wiki/Taxonomy
Not at all familiar (1)
Slightly familiar (2)
Moderately familiar (3)
Very familiar (4)
Extremely familiar (5)
Q16 Do you know any taxonomy of human actions for Action-based Learning and assessment in virtual training environments?*
Yes (1)
No (2)
If your answer is 'Yes' please name those taxonomies of human actions in virtual worlds here.

Appendix 2: Willis input-output three level model

Appendix 3: Berliner classificatory scheme

Appendix 4: Meister taxonomy

Appendix 5: Bennett's Semantic Classificatory Approach (1971)

Appendix 6: Task representing different ability categories

This material has been removed due to copyright restrictions

Appendix 7: Task and skill classification system

Appendix 8: RATaC taxonomy

Appendix 9: HAR process as applied to VE system development

Appendix 10: Performance Shaping Factor taxonomies

Appendix 11: RATaC task taxonomy

PSF categories (Salmon et al., 2008):

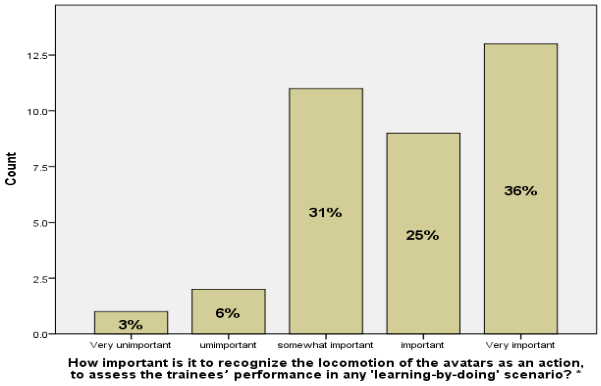
Appendix 12: Maintenance task taxonomy

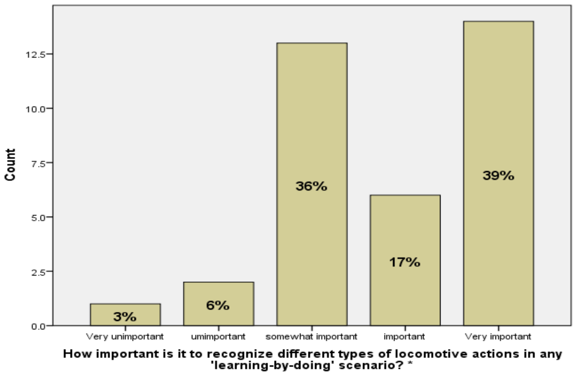
Appendix 13: Karam & Schraefel categories of gestures

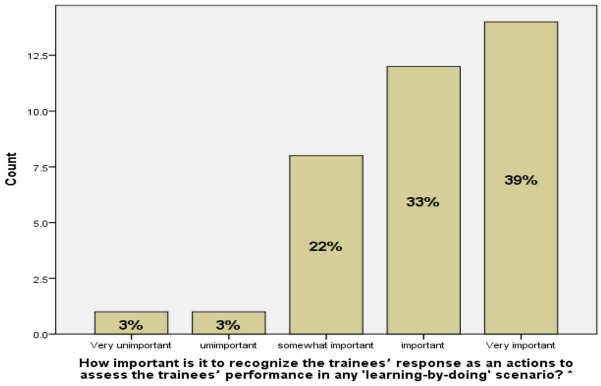
Appendix 14: Data analysis tables and figures

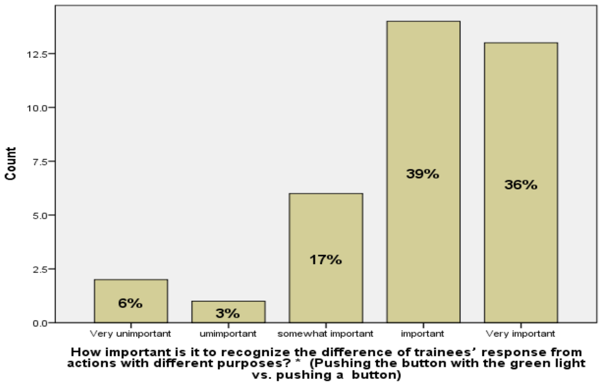
Table 6.9: Descriptive statistics

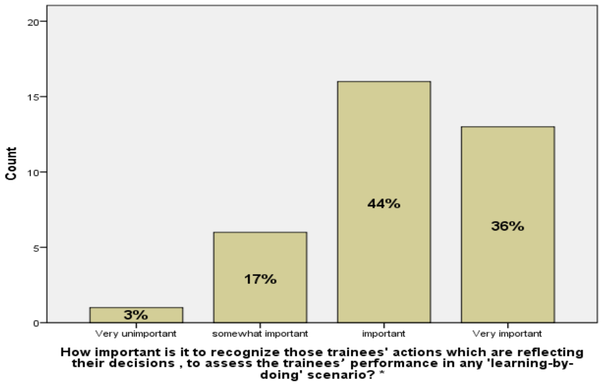
			Observed N	Expected N	Residual
Q4	How important is it to recognise the locomotion of the avatars as an action, to assess the trainees’ performance in any 'Action-based Learning' scenario?	Very unimportant	1	7.2	-6.2
		Unimportant	2	7.2	-5.2
		Somewhat important	11	7.2	3.8
		Important	9	7.2	1.8
		Very important	13	7.2	5.8
Q5	How important is it to recognise different types of locomotive actions in any 'Action-based Learning' scenario?	Very unimportant	1	7.2	-6.2
		Unimportant	2	7.2	-5.2
		Somewhat important	13	7.2	5.8
		Important	6	7.2	-1.2
		Very important	14	7.2	6.8
Q6	How important is it to recognise the trainees’ response as an actions to assess the trainees’ performance in any 'Action-based Learning' scenario? *	Very unimportant	1	7.2	-6.2
		Unimportant	1	7.2	-6.2
		Somewhat important	8	7.2	.8
		Important	12	7.2	4.8
		Very important	14	7.2	6.8
Q7	How important is it to recognise the difference of trainees’ response from actions with different purposes? * (Pushing the button with the green light vs. pushing a button)	Very unimportant	2	7.2	-5.2
		Unimportant	1	7.2	-6.2
		Somewhat important	6	7.2	-1.2
		Important	14	7.2	6.8
		Very important	13	7.2	5.8
Q8	How important is it to recognise those trainees' actions which are reflecting their decisions, to assess the trainees’ performance in any 'Action-based Learning' scenario?	Very unimportant	1	9.0	-8.0
		Unimportant			
		Somewhat important	6	9.0	-3.0
		Important	16	9.0	7.0
		Very important	13	9.0	4.0
Q9	How important is it to recognise the operative actions of trainees to assess their performance in any 'Action-based Learning' scenario?	Very unimportant	1	7.2	-6.2
		Unimportant	2	7.2	-5.2
		Somewhat important	3	7.2	-4.2
		Important	15	7.2	7.8
		Very important	15	7.2	7.8
Q10	How important is it to differentiate the actions which are not changing the objects (like breaking them) or the environment with the actions that are changing the structure or the nature of the objects?	Very unimportant	1	7.2	-6.2
		Unimportant	1	7.2	-6.2
		Somewhat important	8	7.2	.8
		Important	6	7.2	-1.2
		Very important	20	7.2	12.8
Q11	How important is it to recognise the gestures of avatars as a communication method, to assess the trainees’ performance in any 'Action-based Learning' scenario?	Very unimportant	1	7.2	-6.2
		Unimportant	6	7.2	-1.2
		Somewhat important	7	7.2	-.2
		Important	9	7.2	1.8
		Very important	13	7.2	5.8
Q12	How important is it to assess the trainees’ performance based on multiple experts’ solutions instead of one expert?	Very unimportant	2	7.2	-5.2
		Unimportant	2	7.2	-5.2
		Somewhat important	7	7.2	-.2
		Important	12	7.2	4.8
		Very important	13	7.2	5.8
Q13	Most assessments are performed by one assessor, but the Action-based Learning Assessment method uses a panel of assessors instead. What number of experts do you think is more suitable as a panel?	1	2	9.0	-7.0
		1-3	22	9.0	13.0
		3-5	11	9.0	2.0
		5 and more	1	9.0	-8.0
Q14	How important is to have a clear classification of actions in Action-based Learning Assessment?	Very unimportant	1	7.2	-6.2
		Unimportant	2	7.2	-5.2
		Somewhat important	5	7.2	-2.2
		Important	10	7.2	2.8
		Very important	18	7.2	10.8

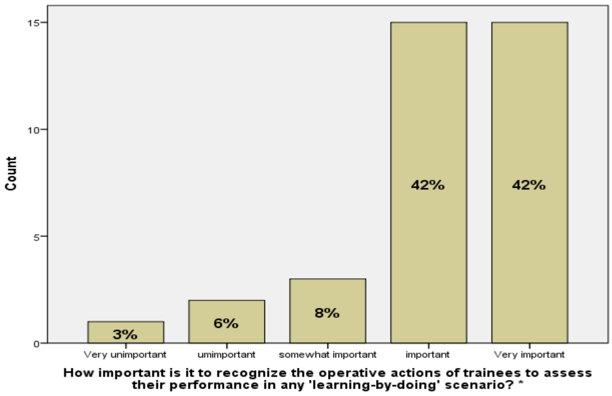

Count
0.0
2.5
5.0
7.5
10.0
12.5
3%
6%
31%
25%
36%
Very unimportant
umimportant
somewhat important
important
Very important
How important is it to recognize the locomotion of the avatars as an action, to assess the trainees' performance in any 'learning-by-doing' scenario? *

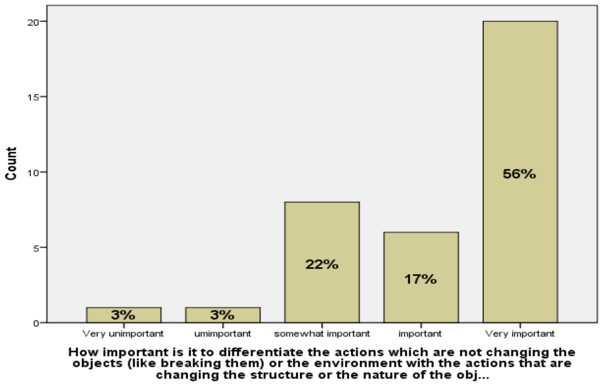

Count
0.0
2.5
5.0
7.5
10.0
12.5
3%
6%
36%
17%
39%
Very unimportant
umimportant
somewhat important
important
Very important
How important is it to recognize different types of locomotive actions in any 'learning-by-doing' scenario? *

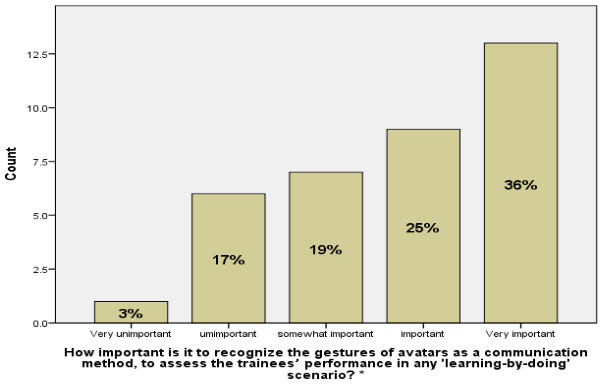

Count
0.0
2.5
5.0
7.5
10.0
12.5
3%
3%
22%
33%
39%
Very unimportant
umimportant
somewhat important
important
Very important
How important is it to recognize the trainees' response as an actions to assess the trainees' performance in any 'learning-by-doing' scenario? *

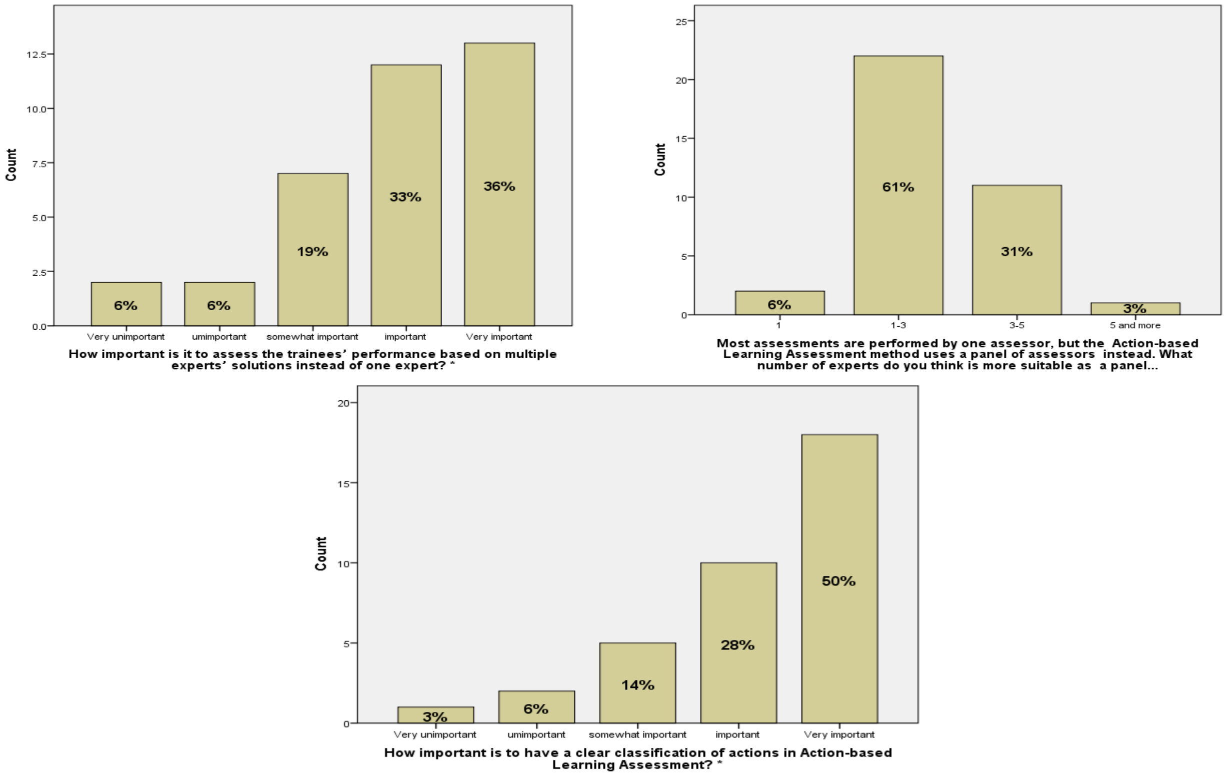

Count
0.0
2.5
5.0
7.5
10.0
12.5
6%
3%
17%
39%
36%
Very unimportant
umimportant
somewhat important
important
Very important
How important is it to recognize the difference of trainees' response from actions with different purposes? * (Pushing the button with the green light vs. pushing a button)


Count
0
5
10
15
20
3%
17%
44%
36%
Very unimportant
somewhat important
important
Very important
How important is it to recognize those trainees' actions which are reflecting their decisions , to assess the trainees' performance in any 'learning-by-doing' scenario? *


Count
0
5
10
15
3%
6%
8%
42%
42%
Very unimportant
umimportant
somewhat important
important
Very important
How important is it to recognize the operative actions of trainees to assess their performance in any 'learning-by-doing' scenario? *


Count
0
5
10
15
20
3%
3%
22%
17%
56%
Very unimportant
umimportant
somewhat important
important
Very important
How important is it to differentiate the actions which are not changing the objects (like breaking them) or the environment with the actions that are changing the structure or the nature of the obj...


Count
0.0
2.5
5.0
7.5
10.0
12.5
3%
17%
19%
25%
36%
Very unimportant
umimportant
somewhat important
important
Very important
How important is it to recognize the gestures of avatars as a communication method, to assess the trainees' performance in any 'learning-by-doing' scenario? *



Figures 6.4 – 6.14: Participants' responses to questions 4 to 14.

Table 6.12: Survey questions' variance

Question number	Question variance
Q1	0.732
Q2	1.467
Q3	1.294
Q4	1.324
Q5	1.32
Q6	1.114
Q7	1.676
Q8	1.206
Q9	1.359
Q10	1.412
Q11	1.477
Q12	1.294
Q13	0.353
Q14	1.676
Q15	1.176
Q16	0.183

Table 6.23: Survey questions' variance

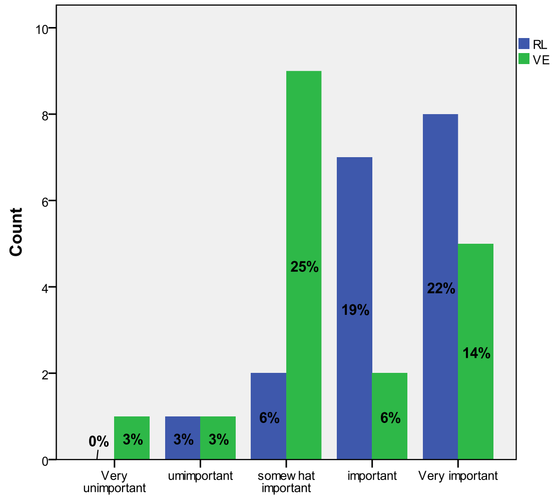
Question number	Question variance
Q1	0.353
Q2	0.588
Q3	0.588
Q4	0.771
Q5	1.046
Q6	0.928
Q7	0.693
Q8	0.252
Q9	0.382
Q10	0.33
Q11	0.918
Q12	1.281
Q13	0.448
Q14	0.382
Q15	0.722
Q16	0

Table 6.30: Descriptive statistics of industry experts

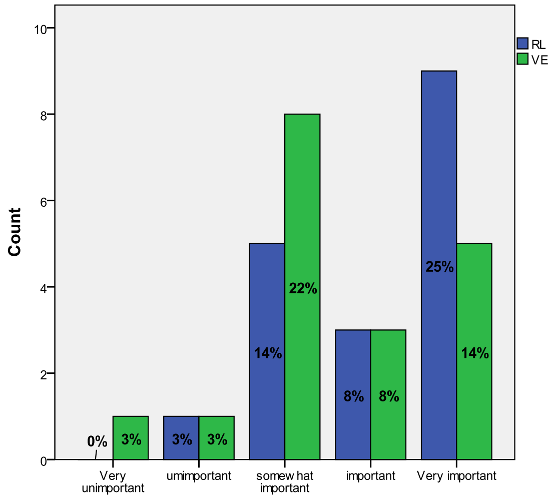
			Observed N	Expected N	Residual
Q4	How important is it to recognise the locomotion of the avatars as an action, to assess the trainees' performance in any 'Action-based Learning' scenario?	Very unimportant			
		Unimportant	1	4.5	-3.5
		Somewhat important	2	4.5	-2.5
		Important	7	4.5	2.5
		Very important	8	4.5	3.5
Q5	How important is it to recognise different types of locomotive actions in any 'Action-based Learning' scenario?	Very unimportant			
		Unimportant	1	4.5	-3.5
		Somewhat important	5	4.5	.5
		Important	3	4.5	-1.5
		Very important	9	4.5	4.5
Q6	How important is it to recognise the trainees' response as an actions to assess the trainees' performance in any 'Action-based Learning' scenario? *	Very unimportant			
		Unimportant	1	4.5	-3.5
		Somewhat important	4	4.5	-.5
		Important	5	4.5	.5
		Very important	8	4.5	3.5
Q7	How important is it to recognise the difference of trainees' response from actions with different purposes? * (Pushing the button with the green light vs. pushing a button)	Very unimportant			
		Unimportant	1	4.5	-3.5
		Somewhat important	2	4.5	-2.5
		Important	9	4.5	4.5
		Very important	6	4.5	1.5
Q8	How important is it to recognise those trainees' actions which are reflecting their decisions, to assess the trainees' performance in any 'Action-based Learning' scenario?	Very unimportant			
		Unimportant			
		Somewhat important			
		Important	11	9.0	2.0
		Very important	7	9.0	-2.0
Q9	How important is it to recognise the operative actions of trainees to assess their performance in any 'Action-based Learning' scenario?	Very unimportant			
		Unimportant			
		Somewhat important	1	6.0	-5.0
		Important	7	6.0	1.0
		Very important	10	6.0	4.0
Q10	How important is it to differentiate the actions which are not changing the objects (like breaking them) or the environment with the actions that are changing the structure or the nature of the objecccts?	Very unimportant			
		Unimportant			
		Somewhat important	1	6.0	-5.0
		Important	3	6.0	-3.0
		Very important	14	6.0	8.0
Q11	How important is it to recognise the gestures of avatars as a communication method, to assess the trainees' performance in any 'Action-based Learning' scenario?	Very unimportant			
		Unimportant	1	4.5	-3.5
		Somewhat important	3	4.5	-1.5
		Important	4	4.5	-.5
		Very important	10	4.5	5.5
Q12	How important is it to assess the trainees' performance based on multiple experts' solutions instead of one expert?	Very unimportant	1	4.5	-3.5
		Unimportant			
		Somewhat important	4	4.5	-.5
		Important	4	4.5	-.5
		Very important	9	4.5	4.5
Q13	Most assessments are performed by one assessor, but the Action-based Learning Assessment method uses a panel of assessors instead. What number of experts do you think is more suitable as a panel?	1	1	4.5	-3.5
		1-3	12	4.5	7.5
		3-5	4	4.5	-.5
		5 and more	1	4.5	-3.5
Q14	How important is to have a clear classification of actions in Action-based Learning Assessment?	Very unimportant			
		Unimportant			
		Somewhat important	1	6.0	-5.0
		Important	7	6.0	1.0
		Very important	10	6.0	4.0

Table 6.31: Descriptive statistics of virtual worlds' experts

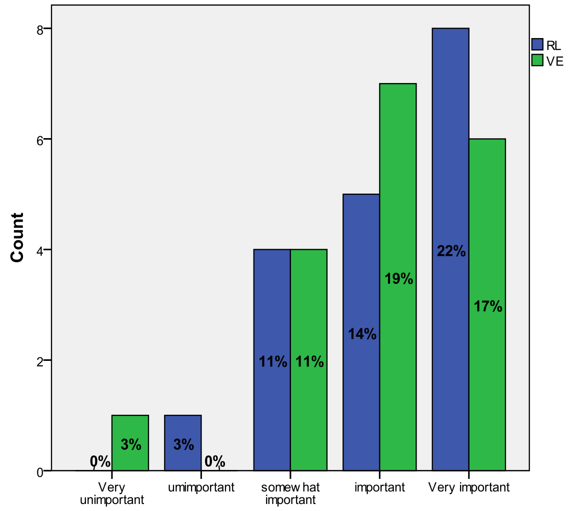
			Observed N	Expected N	Residual
Q4	How important is it to recognise the locomotion of the avatars as an action, to assess the trainees' performance in any 'Action-based Learning' scenario?	Very unimportant	1	3.6	-2.6
		Unimportant	1	3.6	-2.6
		Somewhat important	9	3.6	5.4
		Important	2	3.6	-1.6
		Very important	5	3.6	1.4
Q5	How important is it to recognise different types of locomotive actions in any 'Action-based Learning' scenario?	Very unimportant	1	3.6	-2.6
		Unimportant	1	3.6	-2.6
		Somewhat important	8	3.6	4.4
		Important	3	3.6	-.6
		Very important	5	3.6	1.4
Q6	How important is it to recognise the trainees' response as an actions to assess the trainees' performance in any 'Action-based Learning' scenario? *	Very unimportant	1	4.5	-3.5
		Unimportant			
		Somewhat important	4	4.5	-.5
		Important	7	4.5	2.5
		Very important	6	4.5	1.5
Q7	How important is it to recognise the difference of trainees' response from actions with different purposes? * (Pushing the button with the green light vs. pushing a button)	Very unimportant	2	4.5	-2.5
		Unimportant			
		Somewhat important	4	4.5	-.5
		Important	5	4.5	.5
		Very important	7	4.5	2.5
Q8	How important is it to recognise those trainees' actions which are reflecting their decisions, to assess the trainees' performance in any 'Action-based Learning' scenario?	Very unimportant	1	4.5	-3.5
		Unimportant			
		Somewhat important	6	4.5	1.5
		Important	5	4.5	.5
		Very important	6	4.5	1.5
Q9	How important is it to recognise the operative actions of trainees to assess their performance in any 'Action-based Learning' scenario?	Very unimportant	1	3.6	-2.6
		Unimportant	2	3.6	-1.6
		Somewhat important	2	3.6	-1.6
		Important	8	3.6	4.4
		Very important	5	3.6	1.4
Q10	How important is it to differentiate the actions which are not changing the objects (like breaking them) or the environment with the actions that are changing the structure or the nature of the objects???	Very unimportant	1	3.6	-2.6
		Unimportant	1	3.6	-2.6
		Somewhat important	7	3.6	3.4
		Important	3	3.6	-.6
		Very important	6	3.6	2.4
Q11	How important is it to recognise the gestures of avatars as a communication method, to assess the trainees' performance in any 'Action-based Learning' scenario?	Very unimportant	1	3.6	-2.6
		Unimportant	5	3.6	1.4
		Somewhat important	4	3.6	.4
		Important	5	3.6	1.4
		Very important	3	3.6	-.6
Q12	How important is it to assess the trainees' performance based on multiple experts' solutions instead of one expert?	Very unimportant	1	3.6	-2.6
		Unimportant	2	3.6	-1.6
		Somewhat important	3	3.6	-.6
		Important	8	3.6	4.4
		Very important	4	3.6	.4
Q13	Most assessments are performed by one assessor, but the Action-based Learning Assessment method uses a panel of assessors instead. What number of experts do you think is more suitable as a panel?	1	1	6.0	-5.0
		1-3	10	6.0	4.0
		3-5	7	6.0	1.0
		5 and more			
Q14	How important is to have a clear classification of actions in Action-based Learning Assessment?	Very unimportant	1	3.6	-2.6
		Unimportant	2	3.6	-1.6
		Somewhat important	4	3.6	.4
		Important	3	3.6	-.6
		Very important	8	3.6	4.4



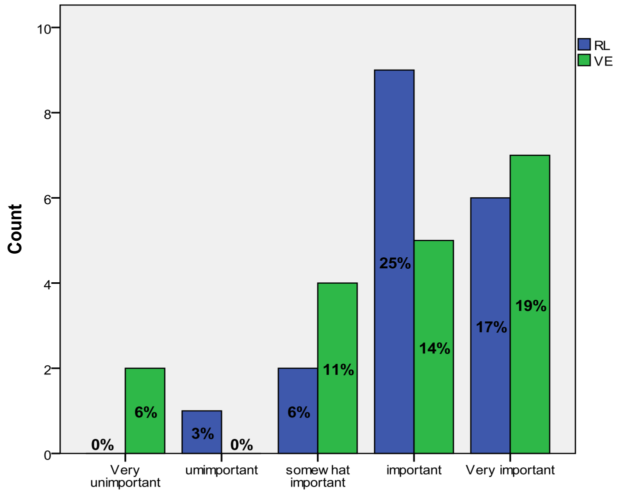
How important is it to recognize the locomotion of the avatars as an action, to assess the trainees' performance in any 'learning-by-doing' scenario? *



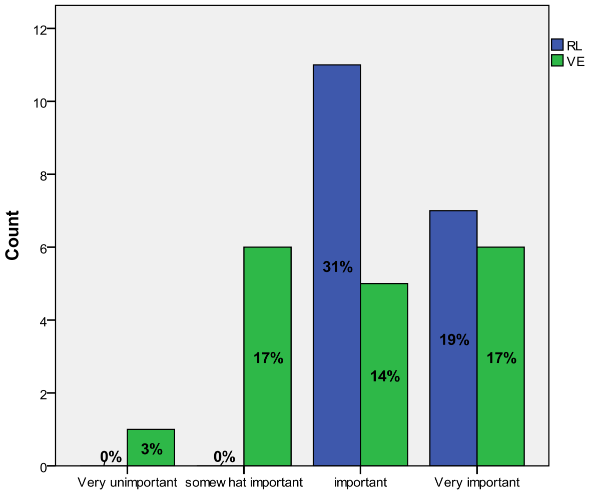
How important is it to recognize different types of locomotive actions in any 'learning-by-doing' scenario? *



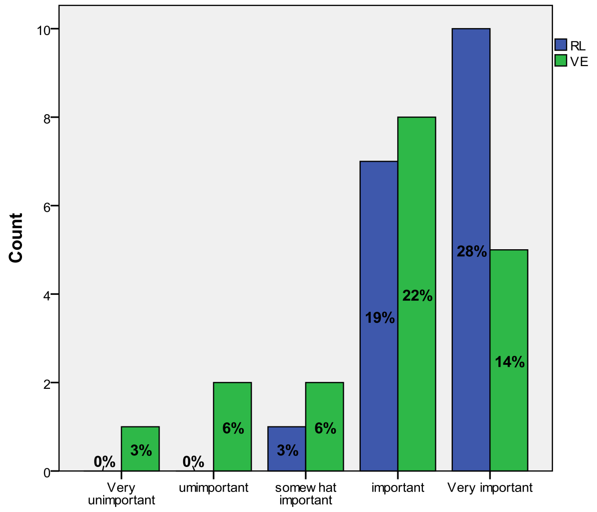
How important is it to recognize the trainees' response as an actions to assess the trainees' performance in any 'learning-by-doing' scenario? *



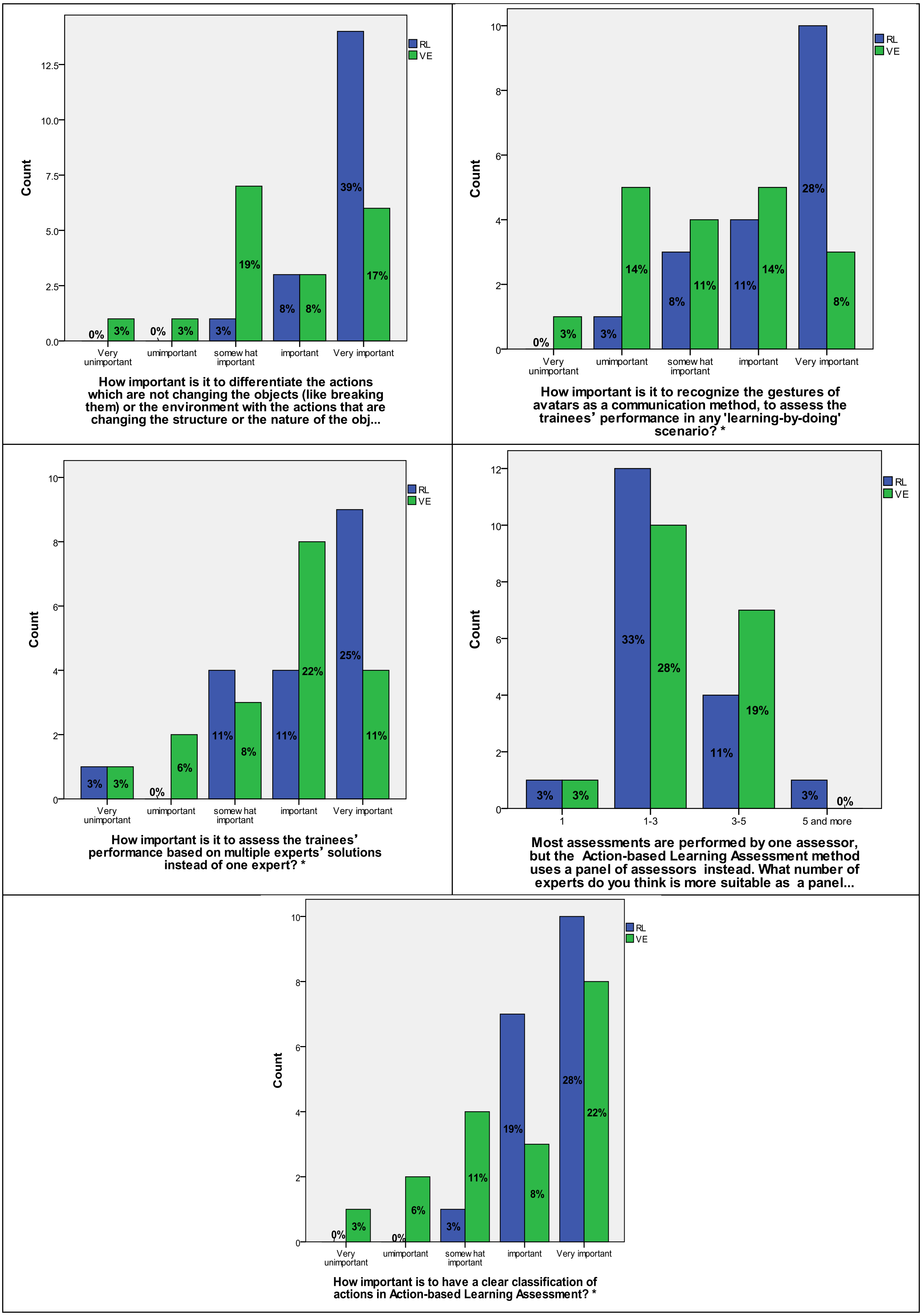
How important is it to recognize the difference of trainees' response from actions with different purposes? * (Pushing the button with the green light vs. pushing a button)



How important is it to recognize those trainees' actions which are reflecting their decisions , to assess the trainees' performance in any 'learning-by-doing' scenario? *



How important is it to recognize the operative actions of trainees to assess their performance in any 'learning-by-doing' scenario? *



How important is it to differentiate the actions which are not changing the objects (like breaking them) or the environment with the actions that are changing the structure or the nature of the obj...

How important is it to recognize the gestures of avatars as a communication method, to assess the trainees' performance in any 'learning-by-doing' scenario? *

How important is it to assess the trainees' performance based on multiple experts' solutions instead of one expert? *

Most assessments are performed by one assessor, but the Action-based Learning Assessment method uses a panel of assessors instead. What number of experts do you think is more suitable as a panel...

How important is to have a clear classification of actions in Action-based Learning Assessment? *

Figures 6.22 – 6.32: Comparing both groups of respondents' views on questions 4 to 14.

Table 6.37: Iteration history and changes in cluster centres

Iteration History[a]						
Iteration	Change in Cluster Centres					
	1	2	3	4	5	6
1	16.855	18.783	30.050	23.445	47.125	12.421
2	2.107	2.087	2.732	2.605	5.891	1.553
3	.263	.232	.248	.289	.736	.194
4	.033	.026	.023	.032	.092	.024
5	.004	.003	.002	.004	.012	.003
6	.001	.000	.000	.000	.001	.000
7	6.430E-5	3.534E-5	1.696E-5	4.412E-5	.000	4.738E-5
8	8.037E-6	3.927E-6	1.542E-6	4.902E-6	2.247E-5	5.923E-6
9	1.005E-6	4.363E-7	1.402E-7	5.446E-7	2.809E-6	7.403E-7
10	1.256E-7	4.848E-8	1.274E-8	6.051E-8	3.511E-7	9.254E-8
11	1.570E-8	5.387E-9	1.159E-9	6.724E-9	4.389E-8	1.157E-8
12	1.962E-9	5.985E-10	1.053E-10	7.471E-10	5.486E-9	1.446E-9
13	2.453E-10	6.651E-11	9.539E-12	8.301E-11	6.858E-10	1.807E-10
14	3.066E-11	7.359E-12	8.773E-13	9.206E-12	8.572E-11	2.259E-11
15	3.832E-12	8.237E-13	9.480E-14	1.024E-12	1.069E-11	2.849E-12
16	4.476E-13	9.804E-14	7.338E-15	1.045E-13	1.367E-12	3.173E-13
17	8.847E-14	2.891E-14	.000	2.963E-14	1.445E-13	6.187E-14
18	3.768E-15	4.965E-16	.000	.000	3.046E-14	1.884E-15
19	8.882E-16	.000	.000	.000	3.580E-15	1.776E-15
20	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centres. The maximum absolute coordinate change for any centre is .000. The current iteration is 20. The minimum distance between initial centres is 235.962.

Table 6.41: Cluster memberships and distances from cluster centres

case	Cluster	Distance	Expert's Clusters
Blench	1	12.81	Responsive
Recede	1	16.18	Responsive
Flinch	1	16.54	Responsive
Recoil	1	19.26	Responsive
Retract	1	21.82	Responsive
Dodge	1	22.26	Responsive
wince	1	28.29	Responsive
Fly	2	9.79	Locomotive
Run	2	11.02	Locomotive
Teleport	2	15.47	Locomotive
Walk	2	15.96	Locomotive
Swim	2	19.62	Locomotive
Crawl	2	21.13	Locomotive
Drive	2	26.09	Locomotive
Jump	2	57.21	Locomotive
Burn	3	25.94	Constructional
Sew	3	26.96	Constructional
Chop	3	28.63	Constructional
Saw	3	29.43	Constructional
Break	3	30.88	Constructional
Smash	3	31.98	Constructional
Cut	3	33.05	Constructional
Drop	3	59.87	Constructional
Drag	3	73.06	Constructional
Throw	3	74.36	Constructional
Close	4	19.35	Operative
Dial	4	19.41	Operative
Carry	4	22.90	Operative
Shut	4	23.60	Operative
Open	4	26.38	Operative
Talk	4	37.47	Operative
Move	4	44.65	Operative
Turn	4	46.89	Operative
Direct	5	11.65	Decisional
Pick	5	13.15	Decisional
Arrange	5	19.32	Decisional
Check	5	24.03	Decisional
Set	5	35.42	Decisional
Collect	5	38.96	Decisional
Choose	5	53.86	Decisional
Wave	6	6.83	Gestural
Sigh	6	12.78	Gestural
Nod	6	13.76	Gestural
Smile	6	14.20	Gestural
Wink	6	16.38	Gestural
Applaud	6	24.84	Gestural
Point	6	32.37	Gestural

Table 6.42: X_{ij} , $\overline{X_J}$ results

Cluster Number	Card	Constructional	Decisional	Gestural	Locomotive	Operative	Responsive
1	wince	3	43	0	1	2	158
1	Dodge	2	2	12	4	6	181
1	Blench	9	23	8	6	5	156
1	Recede	9	6	11	11	10	160
1	Retract	7	3	7	12	19	159
1	Flinch	1	32	2	2	2	168
1	Recoil	3	14	3	2	1	184
Mean ($\bar{x}_j$)		4.86	18.17	6.67	6.00	7.33	163.67
2	Drive	7	1	159	35	4	1
2	Teleport	4	0	188	9	3	3
2	Walk	2	1	190	13	0	1
2	Run	0	1	171	27	1	7
2	Fly	1	0	184	15	4	3
2	Crawl	0	3	194	8	0	2
2	Jump	1	3	133	26	0	44
2	Swim	2	0	193	9	1	2
Mean ($\bar{x}_j$)		2.13	1.13	176.50	17.75	1.63	7.88
3	Drop	108	2	8	56	8	25
3	Throw	97	7	22	68	2	11
3	Drag	103	1	24	73	2	4
3	Break	181	4	4	8	5	5
3	Cut	185	0	7	10	2	3
3	Saw	181	3	5	10	2	6
3	Sew	180	0	6	14	4	3
3	Smash	180	1	6	5	2	13
3	Chop	180	6	5	11	2	3
3	Burn	176	5	3	10	3	10
Mean ($\bar{x}_j$)		157.1	2.9	9	26.5	3.2	8.3
4	Talk	5	30	8	126	14	24
4	Move	9	2	57	131	0	8
4	Turn	3	4	33	114	40	13
4	Shut	19	2	4	164	7	11
4	Carry	25	2	24	155	1	0
4	Dial	7	3	8	163	17	9
4	Close	14	3	5	161	10	14
4	Open	12	0	6	170	14	5
Mean ($\bar{x}_j$)		11.75	5.75	18.125	148	12.875	10.5
5	Choose	2	0	1	6	195	3
5	Pick	6	1	2	33	159	6
5	Arrange	30	1	3	30	138	5
5	Collect	19	0	8	57	121	2
5	Check	5	1	5	20	168	8
5	Set	23	2	5	46	119	12
5	Direct	17	12	2	28	146	2
Mean ($\bar{x}_j$)		14.57	2.43	3.71	31.43	149.43	5.43
6	Point	1	164	3	30	6	3
6	Nod	0	196	0	2	5	4
6	Wink	1	199	1	3	0	3
6	Wave	2	189	4	5	1	6
6	Smile	1	197	0	0	2	7
6	Applaud	2	167	0	6	5	27
6	Sigh	0	183	2	1	0	21
Mean ($\bar{x}_j$)		1.00	185.00	1.43	6.71	2.71	10.14
Mean ($\bar{x}_{ij}$)		38.83	32.32	36.72	40.34	26.77	32.02

Table 6.43: $\left(X_{ij} - \overline{X}_J\right)^2$ results

Cluster Number	Card	Constructional	Decisional	Gestural	Locomotive	Operative	Responsive
1	wince	15870.64	613.3526	1547.669	1348.608	114.0806	1283.774
1	Dodge	22194.66	431.225	1320.627	611.2467	919.2508	1356.433
1	Blench	15370.72	473.7569	1179.265	825.034	86.84654	889.8162
1	Recede	16378.55	281.0973	860.8606	661.6935	692.6976	889.8162
1	Retract	16123.6	60.3101	803.1797	883.4808	859.6125	1013.135
1	Flinch	18490.21	613.3526	1469.988	1205.715	0.101856	1431.093
1	Recoil	23097.53	663.8846	1469.988	1137.268	335.5912	1283.774
2	Drive	962.3196	518.2888	28.52014	14951.57	980.8891	1013.135
2	Teleport	842.2345	564.8207	982.2223	22884.61	1044.527	1213.114
2	Walk	962.3196	716.4165	747.4989	23493.71	980.8891	1356.433
2	Run	626.0643	663.8846	177.967	18030.2	980.8891	1507.752
2	Fly	842.2345	518.2888	642.1372	21690.4	1044.527	1431.093
2	Crawl	901.277	716.4165	1045.903	24735.93	859.6125	1507.752
2	Jump	143.4898	716.4165	205.6478	9269.183	859.6125	1431.093
2	Swim	901.277	663.8846	982.2223	24422.37	1044.527	1356.433
3	Drop	49.29833	352.1612	245.2223	825.034	919.2508	4784.518
3	Throw	441.8941	613.3526	765.0521	216.7786	641.0593	3383.774
3	Drag	785.1919	613.3526	1066.648	161.885	980.8891	4117.816
3	Break	730.1494	473.7569	1045.903	1070.821	801.9742	20212.37
3	Cut	842.2345	613.3526	920.5414	883.4808	1044.527	21365.73
3	Saw	677.1068	613.3526	920.5414	1006.374	859.6125	20212.37
3	Sew	842.2345	518.2888	693.818	943.9276	1044.527	19929.03
3	Smash	361.809	613.3526	1248.946	943.9276	980.8891	19929.03
3	Chop	842.2345	613.3526	860.8606	1006.374	692.6976	19929.03
3	Burn	484.9366	564.8207	920.5414	1137.268	746.3359	18815.67
4	Talk	64.34088	162.9697	7337.563	825.034	5.378452	1144.455
4	Move	577.0217	716.4165	8219.158	411.1403	919.2508	889.8162
4	Turn	361.809	175.1399	5425.733	13.86374	801.9742	1283.774
4	Shut	441.8941	390.6931	15291.69	1070.821	919.2508	393.2205
4	Carry	1025.362	663.8846	13146.82	161.885	919.2508	191.263
4	Dial	529.9792	95.37392	15045.37	825.034	859.6125	1013.135
4	Close	324.7664	281.0973	14558.73	1006.374	859.6125	616.5183
4	Open	730.1494	162.9697	16811.61	943.9276	1044.527	719.8375
5	Choose	842.2345	28302.69	1179.265	1276.162	1044.527	1356.433
5	Pick	677.1068	17485.84	53.88185	1205.715	980.8891	1077.795
5	Arrange	730.1494	12373.01	106.9244	1137.268	980.8891	77.96514
5	Collect	901.277	8880.055	277.5414	825.034	1044.527	393.2205
5	Check	577.0217	19947.05	413.7329	1006.374	980.8891	1144.455
5	Set	400.8515	8507.119	32.03078	1006.374	919.2508	250.5822
5	Direct	901.277	14216.76	152.2861	1205.715	412.8678	476.5396
6	Point	842.2345	431.225	106.9244	1137.268	17339.85	1431.093
6	Nod	785.1919	473.7569	1469.988	1348.608	26791.42	1507.752
6	Wink	842.2345	716.4165	1394.307	1276.162	27782.51	1431.093
6	Wave	677.1068	663.8846	1248.946	1070.821	24548.89	1356.433
6	Smile	626.0643	613.3526	1627.35	1348.608	27119.78	1431.093
6	Applaud	25.21322	473.7569	1179.265	1348.608	18138.93	1356.433
6	Sigh	121.4685	716.4165	1547.669	1205.715	22704.72	1507.752

Table 6.44: $\left(X_{ij}-\overline{X_{\imath\jmath}}\right)^2$ results

Cluster Number	Card	Constructional	Decisional	Gestural	Locomotive	Operative	Responsive
1	wince	32.11111	28.44444	25	44.44444	616.6944	3.44898
1	Dodge	300.4444	1.777778	4	28.44444	261.3611	8.163265
1	Blench	58.77778	5.444444	0	1.777778	23.36111	17.16327
1	Recede	13.44444	7.111111	25	18.77778	148.0278	17.16327
1	Retract	21.77778	136.1111	36	0.111111	230.0278	4.591837
1	Flinch	18.77778	28.44444	16	21.77778	191.3611	14.87755
1	Recoil	413.4444	40.11111	16	13.44444	17.36111	3.44898
2	Drive	47.26563	5.640625	297.5625	306.25	0.015625	23.76563
2	Teleport	23.76563	1.890625	76.5625	132.25	1.265625	3.515625
2	Walk	47.26563	2.640625	22.5625	182.25	0.015625	0.015625
2	Run	0.765625	0.390625	85.5625	30.25	0.015625	4.515625
2	Fly	23.76563	5.640625	7.5625	56.25	1.265625	1.265625
2	Crawl	34.51563	2.640625	95.0625	306.25	3.515625	4.515625
2	Jump	1305.016	2.640625	68.0625	1892.25	3.515625	1.265625
2	Swim	34.51563	0.390625	76.5625	272.25	1.265625	0.015625
3	Drop	278.89	23.04	870.25	1	0.81	2410.81
3	Throw	7.29	1.44	1722.25	169	16.81	3612.01
3	Drag	18.49	1.44	2162.25	225	3.61	2926.81
3	Break	10.89	3.24	342.25	25	1.21	571.21
3	Cut	28.09	1.44	272.25	4	8.41	778.41
3	Saw	5.29	1.44	272.25	16	0.01	571.21
3	Sew	28.09	0.64	156.25	9	8.41	524.41
3	Smash	22.09	1.44	462.25	9	3.61	524.41
3	Chop	28.09	1.44	240.25	16	9.61	524.41
3	Burn	2.89	0.04	272.25	36	4.41	357.21
4	Talk	182.25	1.265625	484	102.5156	588.0625	45.5625
4	Move	6.25	165.7656	289	1511.266	14.0625	7.5625
4	Turn	6.25	735.7656	1156	221.2656	3.0625	76.5625
4	Shut	0.25	34.51563	256	199.5156	14.0625	52.5625
4	Carry	110.25	141.0156	49	34.51563	14.0625	175.5625
4	Dial	2.25	17.01563	225	102.5156	7.5625	22.5625
4	Close	12.25	8.265625	169	172.2656	7.5625	5.0625
4	Open	30.25	1.265625	484	147.0156	33.0625	0.0625
5	Choose	5.897959	2076.755	646.6122	7.367347	5.897959	158.0408
5	Pick	0.326531	91.61224	2.469388	2.938776	2.040816	73.46939
5	Arrange	0.183673	130.6122	2.040816	0.510204	2.040816	238.0408
5	Collect	11.7551	808.1837	653.898	18.36735	5.897959	19.61224
5	Check	6.612245	344.898	130.6122	1.653061	2.040816	91.61224
5	Set	43.18367	925.898	212.3265	1.653061	0.183673	71.04082
5	Direct	11.7551	11.7551	11.7551	2.938776	91.61224	5.897959
6	Point	51.02041	10.79592	542.2245	2.469388	441	0
6	Nod	37.73469	5.22449	22.22449	2.040816	121	1
6	Wink	51.02041	7.367347	13.79592	0.183673	196	0
6	Wave	17.16327	2.938776	2.938776	6.612245	16	1
6	Smile	9.877551	0.510204	45.08163	2.040816	144	0
6	Applaud	284.1633	5.22449	0.510204	2.040816	324	1
6	Sigh	117.8776	7.367347	32.65306	0.326531	4	1

Table 6.45: Percentage of each clustered item

	Constructional	Gestural	Locomotive	Operative	Decisional	Responsive
Cut	89%	0%	3%	5%	1%	1%
Break	87%	2%	2%	4%	2%	2%
Sew	87%	0%	3%	7%	2%	1%
Saw	87%	1%	2%	5%	1%	3%
Chop	87%	3%	2%	5%	1%	1%
Smash	87%	0%	3%	2%	1%	6%
Burn	85%	2%	1%	5%	1%	5%
Drop	52%	1%	4%	27%	4%	12%
Drag	50%	0%	12%	35%	1%	2%
Throw	47%	3%	11%	33%	1%	5%
Wink	0%	96%	0%	1%	0%	1%
Nod	0%	95%	0%	1%	2%	2%
Smile	0%	95%	0%	0%	1%	3%
Wave	1%	91%	2%	2%	0%	3%
Sigh	0%	88%	1%	0%	0%	10%
Applaud	1%	81%	0%	3%	2%	13%
Point	0%	79%	1%	14%	3%	1%
Crawl	0%	1%	94%	4%	0%	1%
Swim	1%	0%	93%	4%	0%	1%
Walk	1%	0%	92%	6%	0%	0%
Teleport	2%	0%	91%	4%	1%	1%
Fly	0%	0%	89%	7%	2%	1%
Run	0%	0%	83%	13%	0%	3%
Drive	3%	0%	77%	17%	2%	0%
Jump	0%	1%	64%	13%	0%	21%
Open	6%	0%	3%	82%	7%	2%
Shut	9%	1%	2%	79%	3%	5%
Dial	3%	1%	4%	79%	8%	4%
Close	7%	1%	2%	78%	5%	7%
Carry	12%	1%	12%	75%	0%	0%
Move	4%	1%	28%	63%	0%	4%
Talk	2%	14%	4%	61%	7%	12%
Turn	1%	2%	16%	55%	19%	6%
Choose	1%	0%	0%	3%	94%	1%
Check	2%	0%	2%	10%	81%	4%
Pick	3%	0%	1%	16%	77%	3%
Direct	8%	6%	1%	14%	71%	1%
Arrange	14%	0%	1%	14%	67%	2%
Collect	9%	0%	4%	28%	58%	1%
Set	11%	1%	2%	22%	57%	6%
Recoil	1%	7%	1%	1%	0%	89%
Dodge	1%	1%	6%	2%	3%	87%
Flinch	0%	15%	1%	1%	1%	81%
Retract	3%	1%	3%	6%	9%	77%
Recede	4%	3%	5%	5%	5%	77%
wince	1%	21%	0%	0%	1%	76%
Blench	4%	11%	4%	3%	2%	75%

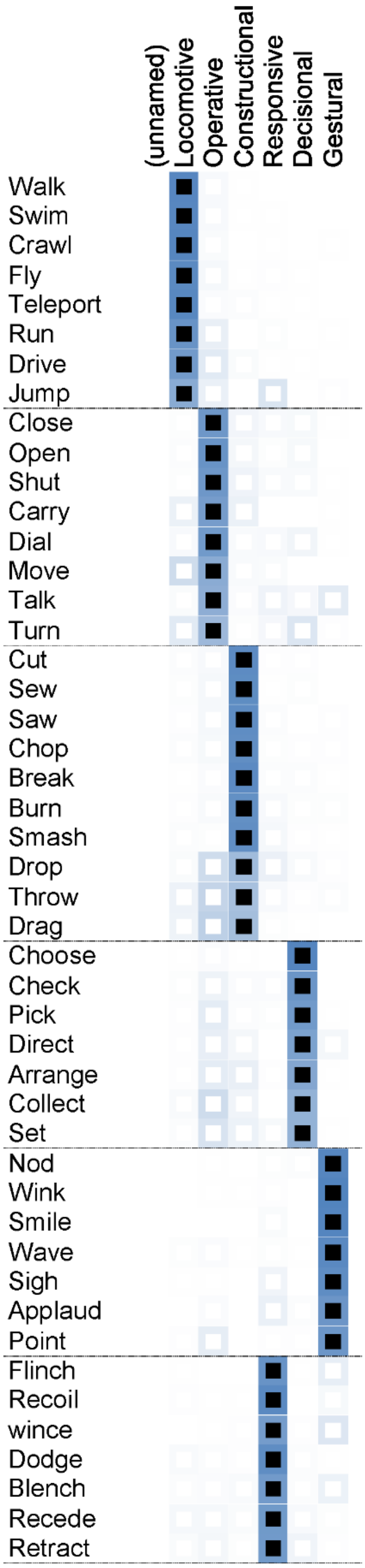


Figure 6.36: Sorted items colour map

Appendix 15: Copyright Documents

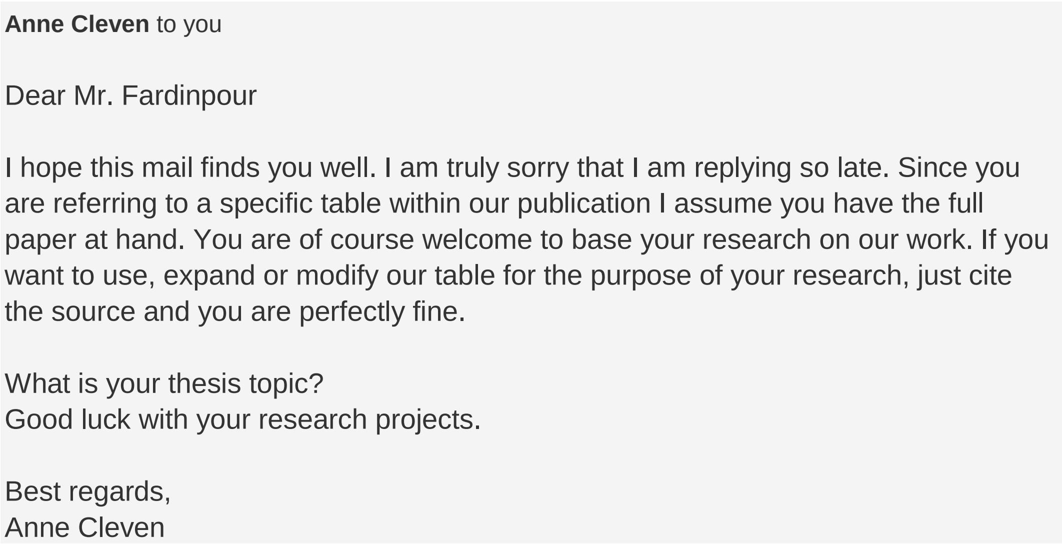
Dear Dr. Cleven
I was informed that to re-use a table from your publication in my PhD thesis I have to ask permission from you as the author. The paper I am interested in is: Cleven, A., Gubler, P., & Hüner, K. M. (2009, May). Design alternatives for the evaluation of design science research artifacts. In Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology (p. 19). ACM.

I was wondering if I may use 'Table 1' of this paper in my thesis with your permission.

Best regards

--
Ali Fardinpour
PhD Student

Information Systems School
Curtin University



Anne Cleven to you

Dear Mr. Fardinpour

I hope this mail finds you well. I am truly sorry that I am replying so late. Since you are referring to a specific table within our publication I assume you have the full paper at hand. You are of course welcome to base your research on our work. If you want to use, expand or modify our table for the purpose of your research, just cite the source and you are perfectly fine.

What is your thesis topic?
Good luck with your research projects.

Best regards,
Anne Cleven

Ali Fardinpour <afardinpour@gmail.com>

Re: Re-use permission

1 message

Ken Peffers <k@peffers.com> Tue, Dec 15, 2015 at 9:32 PM
To: Ali Fardinpour <ali.fardinpour@postgrad.curtin.edu.au>

Dear Mr. Fardinpour

You do have my permission to use that figure.

You might have noticed that the DSRM is also explicated in a paper published in JMIS, winter 2007-8.

Regards

Ken

Ken Peffers, Professor of MIS
Lee Business School, UNLV, 4505 Maryland Pkwy, Las Vegas NV 89154-6034.

Tel 702 807 1181 Skype kgpeffers

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

This Agreement between Ali Fardinpour ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	3780130823336
License date	Jan 01, 2016
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	British Journal of Educational Technology
Licensed Content Title	A framework for monitoring instructional environments in a virtual world
Licensed Content Author	David Chodos,Eleni Stroulia,Sharla King,Mike Carbonaro
Licensed Content Date	Dec 3, 2012
Pages	12
Type of use	Book/Textbook
Requestor type	University/Academic
Is the reuse sponsored by or associated with a pharmaceutical or medical products company?	no
Format	Print and electronic
Portion	Figure/table
Number of figures/tables	1
Original Wiley figure/table number(s)	Table 1
Will you be translating?	No
Title of new book	Taxonomy of Human Actions for Action-based Learning Assessment in Virtual Training Environments
Publisher of new book	Curtin University
None	
Author of new book	Ali Fardinpour
Expected publication date of new book	Jul 2016
Estimated size of new book (pages)	300

Requestor Location	Ali Fardinpour 1/18 Stannard St, ST James Perth, Australia 6102 Attn: Ali Fardinpour
Billing Type	Invoice
Billing Address	Ali Fardinpour 1/18 Stannard St, ST James Perth, Australia 6102 Attn: Ali Fardinpour
Total	0.00 USD

Terms and Conditions

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.

- You are hereby granted a personal, non-exclusive, non-sub licensable (on a standalone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, and any CONTENT (PDF or image file) purchased as part of your order, is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.

- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts, You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.

- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY

SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN
CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION,
WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT,
NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT
LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE,
BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY
OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or
excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.

- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and
  Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-

Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing

purposes requires further explicit permission from Wiley and will be subject to