Abstract: This deliverable is the second version of the GRAPPLE evaluation framework and continues the work presented in the first version. The present document complements the existing evaluation framework and aims at refining and extending the evaluation methodology, particularly expatiating on the summative phase of evaluation. The evaluation methodology in general covers evaluation of technical, usability aspects as well as psycho-pedagogical aspects. As a follow up to the formative evaluation cycle addenda to the respective methodology (as developed in the first versions of D8.1a Evaluation Framework and D8.2a Evaluation Guidelines) are outlined, namely the online survey tools used and adjustments to the methodology according to the requisites of the actual empirical settings of university and corporate users. Methodological issues in view of the summative evaluation phase are presented. These shall serve as a basis for improving and amending the evaluation methodology by incorporating additional relevant scientific literature, experiences from the conducted empirical evaluations, as well as feedback from the second year project review. Grounding on these elaborations the methodology for the summative evaluation of the GRAPPLE services and tools is presented. Different levels and types of evaluation are envisaged, providing a comprehensive consideration of the GRAPPLE system from different perspective. Finally, an outlook to next steps is given.

Keyword list: evaluation, evaluation methodology, formative evaluation, summative evaluation, usability, learning effectiveness, user acceptance, adaptation quality, qualitative methods
Authors

<table>
<thead>
<tr>
<th>Person</th>
<th>Email</th>
<th>Partner code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Christina Steiner</td>
<td><a href="mailto:chr.steiner@uni-graz.at">chr.steiner@uni-graz.at</a></td>
<td>UniGraz</td>
</tr>
<tr>
<td>Eva Hillemann</td>
<td><a href="mailto:eva.hillemann@uni-graz.at">eva.hillemann@uni-graz.at</a></td>
<td>UniGraz</td>
</tr>
</tbody>
</table>

Table of Contents

AUTHORS ......................................................................................................................................................... 2

TABLE OF CONTENTS .................................................................................................................................... 2

TABLES AND FIGURES ................................................................................................................................... 4

LIST OF ACRONYMS AND ABBREVIATIONS ............................................................................................... 5

1 INTRODUCTION ........................................................................................................................................ 6

1.1 Evaluation Work Packages and Tasks ............................................................................................... 6

1.2 Evaluation in GRAPPLE ....................................................................................................................... 6

1.3 Objective of this Document ................................................................................................................. 7

2 ADDENDA TO THE FORMATIVE EVALUATION METHODOLOGY....................................................... 8

2.1 Online Survey Tools ............................................................................................................................. 8

2.1.1 SurveyMonkey .................................................................................................................................. 8

2.1.2 Overview of LimeSurvey .................................................................................................................. 9

2.1.3 Online Questionnaires Used .......................................................................................................... 10

2.2 Adjustments of the Evaluation Methodology .................................................................................. 12

2.2.1 WP9-specific Adjustments ............................................................................................................. 13

2.2.2 WP10-specific Adjustments ........................................................................................................... 14

2.2.3 Adjustments for the Visualisation Evaluation ................................................................................ 15

3 REFLECTION ON THE FORMATIVE EVALUATION METHODOLOGY TOWARDS SUMMATIVE EVALUATION ................................................................................................................................................. 16

4 METHODOLOGICAL ISSUES FOR THE SUMMATIVE EVALUATION PHASE ........................................... 18

4.1 Evaluation Topics and Instruments in Adaptive System Evaluation ............................................ 18

4.2 Interpreting Usability Scores ............................................................................................................. 20

4.2.1 System Usability Scale .................................................................................................................. 20

4.2.2 IsoMetrics ...................................................................................................................................... 21
A1 Think-Aloud Method ................................................................................................................................ 46
  A1.1 The Think-Aloud Method in E-Learning Evaluation ........................................................................ 46
  A1.2 Strengths and Weaknesses of the Think-Aloud Method ................................................................ 46
  A1.3 Data Analysis of Think-Alouds .................................................................................................... 47
B1 Focus Groups ........................................................................................................................................... 47
  B1.1 Focus Groups compared to Surveys and Individual Interviews .................................................. 47
  B1.2 Strengths and Weaknesses of the Focus Group Method ............................................................... 48
  B1.3 Procedure and the Level of Moderator Involvement ................................................................. 48
  B1.4 Data analysis ................................................................................................................................ 49

Tables and Figures

List of Tables

Table 1: Adapted questionnaire versions for the different visualization types. ........................................... 16

List of Figures

Figure 1: Overview of the evaluation tasks and work packages. ................................................................. 6
Figure 2: Overview of the objectives of the evaluation phases. ................................................................. 7
Figure 3: Screenshot of SurveyMonkey. ..................................................................................................... 8
Figure 4: Screenshot of LimeSurvey. ......................................................................................................... 9
Figure 5: Screenshot of the UFQ in SurveyMonkey ................................................................................. 10
Figure 6: Screenshot of the IsoMetrics questionnaire in LimeSurvey ...................................................... 11
Figure 7: Overview of evaluation stages as suggested in D8.2a Evaluation Guidelines. ......................... 13
Figure 8: Evaluation stages covered by WP9 evaluations. ...................................................................... 14
Figure 9: Evaluation stages covered by WP10 evaluations. .................................................................... 15
Figure 10: Design process and associated evaluation goals and phases (adapted from Van Velsen et al., 2008)........................................................................................................................................ 20
Figure 11: Adjective rating scale added to the SUS (adopted from Bangor, Kortum and Miller, 2009, p.117). 21
Figure 12: Mean SUS score ratings corresponding to the seven adjective ratings (adopted from Bangor, 21
  Kortum and Miller, 2009, p119). .................................................................................................................. 21
Figure 13: Example usability profile of IsoMetrics results. ..................................................................... 22
Figure 14: Example usability profiles of IsoMetrics results for two different software versions. ............... 23
Figure 15: Pilot evaluation model of TENCompetence, showing the implementation steps and evaluation 27
  of learning outcomes (adopted from Schoonenboom et al., 2007, p.44) ................................................. 27
Figure 16: Summative evaluation scheme for GRAPPLE ........................................................................ 30
Figure 17: Evaluation study design for a layered evaluation of adaptation processes. ............................ 37
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>APQ</td>
<td>Activity Perception Questionnaire</td>
</tr>
<tr>
<td>CAM</td>
<td>Conceptual adaption model</td>
</tr>
<tr>
<td>CHAT</td>
<td>Codes for the Human Analyse of Transcripts</td>
</tr>
<tr>
<td>D (Dx.x)</td>
<td>Deliverable (referring to other deliverables of the GRAPPLE project, e.g. D8.2)</td>
</tr>
<tr>
<td>DQ</td>
<td>Demographic Questionnaire</td>
</tr>
<tr>
<td>DM</td>
<td>Domain Model</td>
</tr>
<tr>
<td>e.g.</td>
<td>for example (exempli gratia)</td>
</tr>
<tr>
<td>et al.</td>
<td>and others (et alii)</td>
</tr>
<tr>
<td>GALE</td>
<td>GRAPPLE Adaptive Learning Environment</td>
</tr>
<tr>
<td>GAT</td>
<td>Grapple Authoring Tool</td>
</tr>
<tr>
<td>GAT</td>
<td>Gesprächsanalytisches Transkriptionssystem</td>
</tr>
<tr>
<td>GRAPPLE</td>
<td>Generic Responsive Adaptive Personalised Learning Environment</td>
</tr>
<tr>
<td>HIAT</td>
<td>Halb-interpretative Arbeits-Transkription</td>
</tr>
<tr>
<td>i.e.</td>
<td>that is (id est)</td>
</tr>
<tr>
<td>IMI</td>
<td>Intrinsic Motivation Inventory</td>
</tr>
<tr>
<td>LMS</td>
<td>Learning Management System</td>
</tr>
<tr>
<td>LTfLL</td>
<td>Language Technologies for Lifelong Learning</td>
</tr>
<tr>
<td>NASA-TLX</td>
<td>NASA Task Load Index</td>
</tr>
<tr>
<td>p.</td>
<td>page (in quotations)</td>
</tr>
<tr>
<td>QGAT</td>
<td>Questionnaire on Grapple Authoring Tools</td>
</tr>
<tr>
<td>SIQ</td>
<td>Subjective Impression Questionnaire</td>
</tr>
<tr>
<td>SUCA</td>
<td>Questionnaire on the confidence in handling computer and computer applications</td>
</tr>
<tr>
<td>SUS</td>
<td>System Usability Scale</td>
</tr>
<tr>
<td>T (Tx.x)</td>
<td>Task (referring to tasks of the GRAPPLE project, e.g. T8.1)</td>
</tr>
<tr>
<td>UFQ</td>
<td>User Feedback Questionnaire</td>
</tr>
<tr>
<td>UFQL</td>
<td>User Feedback Questionnaire for Learners</td>
</tr>
<tr>
<td>UML</td>
<td>Unified modelling process</td>
</tr>
<tr>
<td>VECA</td>
<td>Questionnaire on the familiarity with different computer applications</td>
</tr>
<tr>
<td>vs.</td>
<td>versus</td>
</tr>
<tr>
<td>WP</td>
<td>Work Package</td>
</tr>
<tr>
<td>w.r.t.</td>
<td>with respect to</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Evaluation Work Packages and Tasks

The work done in WP8 and presented in this document constitutes the theoretical elaboration of the evaluation methodology for GRAPPLE. It is closely interrelated with the work in the evaluation and deployment settings, i.e. university and corporate environments, which serve the empirical investigation of the perception, benefits, and effectiveness of the tools and facilities provided by the GRAPPLE environment with potential future users. An overview on the relation between the tasks of the three work packages involved in evaluation is provided in Figure 1.

WP8 is engaged with the theoretical development and elaboration of an evaluation framework and guidelines. These form the basis for the evaluation of the GRAPPLE environment from a technical as well as from a learning/psycho-pedagogical perspective. In WP9 and WP10 trainings and empirical evaluation studies with end users are conducted in the deployment settings following the framework and guidelines provided by WP8. The experiences made in the first (i.e. formative) empirical evaluation cycle shall feed back into WP8 for the refinement of the evaluation methodology. This refined methodology in turn constitutes the basis for the second (i.e. summative) evaluation cycle in empirical settings.

1.2 Evaluation in GRAPPLE

This document shall extend and refine the framework for evaluation in GRAPPLE. Evaluation of e-learning and of software in general aims at answering one or more of three simple questions (compare e.g. Gediga, Hamborg, & Düntsch, 2002):

- How good is it? → i.e. the perception and benefit of a learning technology
- What can be improved? → i.e. the identification of issues for further improvement of a learning technology
- Which one is better? → i.e. comparing different (versions of) learning technologies

In GRAPPLE evaluation refers to the examination of the tools and services developed in GRAPPLE for personalised presentation and consumption of adaptive content - from the technological as well as psycho-pedagogical perspective. Hereby, formative evaluation as well as summative evaluation is addressed. The first one refers to evaluation aiming at gathering information on the improvement of an e-learning technology and design supporting aspects during development and the latter one means evaluation in terms of the investigation of the final achievements and benefits of a system.
The objective of the second and final version of the GRAPPLE evaluation framework presented in this document is to continue the work presented in the first version (i.e. D8.1a Evaluation Framework). In particular, the evaluation methodology shall be complemented and refined towards the summative evaluation phase. The first evaluation cycle particularly addressed the usability and user acceptance of the GRAPPLE Authoring Tools (compare Figure 2). This is because the evaluation and quality assurance of those tools constitutes a critical precondition for a detailed consideration of adaptive learning experiences. Therefore, the focus of the first versions of the evaluation framework and guidelines was on the evaluation from an authoring perspective. The learners’ point of view, though, was not completely disregarded in the evaluation methodology and the empirical evaluations. Rather, the opportunity of letting authors slip into the role of learners for gathering also initial feedback from a learning perspective was recognised and realised.

1.3 Objective of this Document

This document aims at refining and expanding evaluation framework towards the final evaluation phase, which will investigate and demonstrate the final achievement of GRAPPLE and the benefits of the personalised presentation and consumption of adaptive learning content. It can be seen as a complement to the first version, which elaborated in detail on the general evaluation objectives, topics, and scope in GRAPPLE. Summative evaluation shall address more closely the investigation of adaptive learning, i.e. consider adaptation quality and learning effectiveness of the use of the GRAPPLE adaptive learning environment (see Figure 2). Moreover, the evaluation of the GRAPPLE Authoring Tools will be carried on in order to examine the effects of the improvements of the tools inspired from the formative evaluation results and realised in the final implementation phase. The evaluation framework is extended by incorporating experiences made and needs for improvement identified, as well as by taking into account additional relevant research literature and the recommendations from the project review.

This document is structured as follows: As a follow up to the formative evaluation cycle addenda to the respective methodology (as developed in the first versions of D8.1a Evaluation Framework and D8.2a Evaluation Guidelines) are outlined. On the one hand, this refers to the presentation of the online survey tools and questionnaires applied. On the other hand, this refers to adjustments on the evaluation procedure that have been negotiated and coordinated through continuous dialogue and cooperation with WP9 and WP10 to correspond to the needs and conditions of the deployment setting in question. Subsequently, reflections on the formative evaluation methodology towards summative evaluation are discussed and relevant methodological issues in view of the summative evaluation phase are presented. The aim of these elaborations is to improve and amend the evaluation framework. This involves the elaboration of relevant state of the art under consideration of experiences from the practical work with end users and of the feedback from the second year project review. On the basis of this, the summative evaluation methodology is further developed, outlining the general objectives and procedure for the evaluation from an authoring perspective as well as from a learning perspective. In addition, the methodology for evaluating the visualizations of domain and user model data is sketched. In the summative phase this strand of evaluation aims at targeting the benefit and perception of the implemented visualization widgets. Finally, an outlook to upcoming further work, in particular referring to the evaluation guidelines, is presented.
2 Addenda to the Formative Evaluation Methodology

2.1 Online Survey Tools

For a smooth evaluation process in the evaluation events of WP 9 and WP 10, the evaluation questionnaires adopted and developed in the scope of the D8.2a Evaluation Guidelines Document have been prepared in form of online question-and-answer surveys. In this way data collection and recording could be enhanced and facilitated. Two different online survey tools were used: ‘SurveyMonkey’, a free online survey software and questionnaire tool, and the survey software ‘LimeSurvey’ provided via the University of Graz to its members for administering and publishing surveys. In the following a short overview on these survey tools shall be given, to provide give an idea about the facilities of this tool and the look and feel of the questionnaires used.

2.1.1 SurveyMonkey

SurveyMonkey (http://www.surveymonkey.com/) is an online tool for creating and publishing Web-based surveys. The basic functionalities are freely available (i.e. basic account); there is also an enhanced paid product with additional features and services available. The tool provides the opportunity to easily create online surveys and to collect and analyse responses (see Figure 3 for a screenshot).

New surveys can be created from scratch by using a survey template or by copying an existing survey. For adding questions to a survey it can be chosen between different types (e.g. multiple choice, single textbox) and features (e.g. adding a comment field, using the require answer feature or incorporating text validation). The presentation and design of the survey can be customised. Having finished a survey it can be published to collect responses by sending out an email invitation, by posting a link on a website, or by using one’s own email client. For the collection of responses different settings and restrictions can be selected (e.g. multiple responses, editing of responses etc.). For the data gathered a response summary can be retrieved, responses can be filtered, individual responses can be browsed, and the data can be exported in spreadsheet format.

We have used SurveyMonkey for GRAPPLE purposes in order to conduct a first and short online user survey on the GRAPPLE Authoring Tools (compare D9.3 evaluation report, section 4 ‘Results of the first user survey on GAT’). The tool was chosen as the easiest and quickest solution for creating the envisaged short survey. To this end, a basic account was created and used. As the basic subscription, however, does only allow a limited number of questions per survey and does not support a direct export of the gathered evaluation data, for the more comprehensive surveys an alternative tool was used (see the following section on LimeSurvey).

![SurveyMonkey Screenshot](image)

*Figure 3: Screenshot of SurveyMonkey.*
2.1.2 Overview of LimeSurvey

LimeSurvey (http://www.limesurvey.org/) is an open source online survey application and, similar to SurveyMonkey, it enables users to develop and publish surveys without coding knowledge (see Figure 4 for a screenshot). This tool was not used exclusively and from the beginning on (instead of SurveyMonkey) as it was identified as a possible and valuable opportunity only after setting up the online survey for demonstration evaluations. In addition, initial technical support from the administrative side of the university was needed in order to properly use this technology, which did allow applying LimeSurvey only for the evaluation and training events but not for the earlier demonstration evaluations. It allows users to quickly create intuitive and powerful online question-and-answer surveys that can work for many participants without much effort. It supports and facilitates data collection and analysis by dumping all the results into spreadsheets. So it is not necessary to insert each item one by one in a calculation file, it works automatically and makes the work more economical and easier. Furthermore it is possible to export the data to SPSS.

LimeSurvey is in general a more sophisticated tool allowing even greater flexibility and functionality for the creation and administration of online surveys than SurveyMonkey. A survey in LimeSurvey consists of three basic elements: a survey name, at least one group, at least one question.

First of all it is necessary to specify the survey by getting the unique title provided by the survey name. A new survey can be created from scratch, by using a stored survey, or by importing an existing one. Welcome message and description of the survey, presentation and navigation settings, publication and access control, and notification and data management of the survey can be specified and edited. Each question of a survey is member of a question group. The question group allows the user to define logical sections, common subjects themes, or simply facilitates managing a great number of questions.

The questions are the most important part of the survey. There is no limit for the number of questions or question groups in a survey. A question includes the actual question text and the settings where different question types can be selected. Furthermore, it is possible to specify a help text for each question and to define whether the question is mandatory or optional.

Before activating a survey it can be tested at any point while creating it. This allows checking how the survey looks and if it works properly before actually publishing it. After activating the survey and collecting data, LimeSurvey allows viewing the results, all or individual responses can be browsed. It is also possible to generate graphs of the results. For statistical computations the gathered data can be exported to other applications such as Microsoft Word, Microsoft Excel, CSV File, or PDF. It is also possible to export the gathered data to SPSS for further statistical analyses.

Figure 4: Screenshot of LimeSurvey.
2.1.3 Online Questionnaires Used

In almost all evaluations online questionnaires have been used. In addition to their online version, all questionnaires were prepared and provided also as print-versions in case of unavailability of or problems with computer and/or internet connection. (Please note that details on these evaluation instruments can be found in the D8.2a Evaluation Guidelines document and further information on the conduction and results of the survey are reported in the respective evaluation reports of WP9 (i.e. D9.3) and WP10 (i.e. D10.3)).

For the first survey on GAT a questionnaire was created and administered with the SurveyMonkey tool (see Figure 5 for a screenshot; details and results of this survey have been reported in D9.3). The questionnaire was adopted from the D8.2a Evaluation Guidelines document and consisted of three parts, which were administered on separate pages in the sequence as presented below. The main part consisted of an adapted version of the System Usability Scale (SUS; Brooke, 1996) for gathering a general subjective assessment of usability (User Feedback Questionnaire – UFQ). The SUS is commonly used for gathering a quick and general usability evaluation after the respondent had the opportunity to see and use the system. It is a short and easy to administer questionnaire consisting of 10 items that are answered on a 5-point rating scale ranging from ‘strongly disagree’ to ‘strongly agree’. To enrich the data collection by qualitative data, the questionnaire was complemented by three open questions (UFQQ) on the things/aspects liked best, least, and suggested to be improved for the tools. In addition, demographic and background information (gender, age, profession, and experience with virtual learning systems) was queried with a short demographic questionnaire (DQ). The online survey concluded with an end message saying thanks and citing the URL to the GRAPPLE homepage for further information on the project.

![User Feedback Questionnaire (UFQ)](image)

Figure 5: Screenshot of the UFQ in SurveyMonkey.

For the more in-depth evaluation workshops conducted in WP9 and WP10 online questionnaires were created and administered by the use of the LimeSurvey tool and facilities (see Figure 6 for an example screenshot). In case of WP9 two separate surveys were created – one for gathering background information about the workshop participants, and another one for gathering detailed feedback on the usability and functionality of the authoring tools. The first survey ‘Background Information’ consisted of the demographic questionnaire (DQ) and two subscales assessing computer literacy (VECA and SUCA) as described in D8.2a. DQ, VECA, and SUCA were presented on three separate pages.
The second survey used in WP9, focusing on the evaluation of the authoring tools, consisted of a collection of three questionnaires (presented on different pages) for a comprehensive data collection on user acceptance, usability and functionality. These questionnaires correspond to the evaluation instruments outlined in evaluation guidelines for the stage 3 training provider evaluations (D8.2a, section 7). First, a questionnaire addressing relevant aspects of user acceptance according to the Technology Acceptance Model (Davis, Bagozzi, & Warshaw, 1989) was administered. The ‘Subjective Impression Questionnaire’ (SIQ) collects data on the subscales perceived usefulness, perceived ease of use, attitude, and intention to use are relevant factors for behavioural intentions and actual use (Jung, Loria, Mostaghel, & Saha, 2008). Subsequently, a questionnaire especially focusing on GAT and its functionality was used – the Questionnaire on GRAPPLE Authoring Tools (QGAT). The questionnaire consists of in total 32 statements to be answered on a rating scale and the request/opton to provide additional qualitative feedback on each question. The items are grouped into subscales on the individual authoring tools (DM tool, CRT tool, CAM tool, simulation authoring tool, virtual reality authoring tool) and one subscale on GAT as a whole. For the evaluation workshops only items on GAT in general and the three main tools were used and presented in the questionnaire (i.e. items 1-24). Finally, the IsoMetrics usability inventory (Gediga, Hamborg, & Düntsch, 1999) was applied as a standard usability instrument for gathering a general and comparable usability result (see Figure 6). It consists of 75 items that are leading to scores on seven different aspects/principles of usability (suitability for the task, self descriptiveness, controllability, conformity with user expectations, error tolerance, suitability for individualisation, and suitability for learning).

The same two surveys were also used in the WP10 training and evaluation workshops. In WP10, though, a third survey was used. In those workshops also a short session of letting authors browse through an example adaptive course and thus, simulate a learning experience, was realised. The additional survey

![IsoMetrics](image)

**Figure 6: Screenshot of the IsoMetrics questionnaire in LimeSurvey.**
contained the evaluation instruments as presented in the evaluation guidelines for stage 2 learner evaluations (D8.2a, section 6) – each instrument was presented on a separate page in the survey. Usability was measured using the Short Usability Scale (Brooke, 1996), complemented by additional questions gathering qualitative usability data, data on user acceptance, as well as feedback on the perceived adaptation quality. Altogether, this formed the User Feedback Questionnaire for Learners (UFQL). Subjective cognitive effort was assessed with the ‘Workload Perception Questionnaire’ (WPQ), which actually covers the NASA Task Load Index (Hart & Staveland, 1988). Finally, two subscales of the Intrinsic Motivation Inventory (IMI; Ryan, 1982) were administered in the Activity Perception Questionnaire (APQ); the subscale ‘Interest/Enjoyment’ as an indicator for user acceptance and the subscale ‘Perceived Competence’ as a subjective learning effectiveness measure.

For the evaluation of the mock-up visualisations for domain and learner model data created in WP4 (T4.5) on the basis of the evaluation guidelines (D8.2a, section 8) a questionnaire was used for investigating the perceived benefits, usability, and acceptance of the visualizations. In the beginning a few questions on demography were posed. As the different visualisations are intended either for instructors or for learners, it was decided that a respondent should only be presented with those visualisations corresponding to his/her role in the teaching/learning process and provide feedback on them. Correspondingly, three different branches or versions were realised through the application of a conditional question. Correspondingly, a respondent was presented either with only the visualisations for learners or for instructors or both.

2.2 Adjustments of the Evaluation Methodology

In the course of concretely planning the conduction of empirical evaluations in GRAPPLE’s deployment settings a few adjustments of the evaluation methodology as proposed by the evaluation guidelines document (D8.2a) became necessary to ensure a smooth evaluation process and to come up to the requisites of the empirical settings.

One issue that constituted a slight adjustment of the evaluation methodology, more concretely, of one the instruments used, refers to the IsoMetrics standard usability scale (Gediga et al., 1999) that was applied for the usability evaluation of the authoring tools. Originally, it was planned to apply the formative version of the IsoMetrics (i.e. IsoMetricsL), which in addition to gathering data/scores on seven different aspects of usability also gathers qualitative data for each of the 75 items. In the course of planning and preparing the training/evaluation events it turned out, however, that it would be unfeasible to use this version of the questionnaire (even if using only some subscales). The estimated duration (according to the manual) for filling out the complete long version of the questionnaire was several hours and thus, exceeded the total time planned for the training and evaluation events. Besides, requesting participants to fill out a questionnaire taking that much time was assumed to be rather unreasonable. Participants were assumed to likely show reactance, demotivation, reluctance, and possibly unwillingness to participate/proceed. Therefore it was decided to use in both WP9 and WP10 only the short version of the questionnaire (IsoMetricsS), which is recommended for summative evaluation. This instrument provided a standardised and comparable measure of usability - yielding general scores for the different usability aspects but no in-depth, qualitative data on issues for improvement. It was assumed that important and helpful qualitative feedback and even more concrete and tangible information on needs for improvement could be better gathered through the Questionnaire on GRAPPLE Authoring Tools (QGAT). QGAT was decided to be presented to participants before the IsoMetrics. In this way, it should be ensured that people were not yet too tired to appropriately respond – especially as the IsoMetrics questionnaire, even in its short version, might be perceived rather long.

Further adaptations of the evaluation methodology have been made with respect to the realisation of the three stages of training and evaluation. As the training on the GRAPPLE facilities actually is necessary for the empirical evaluation, evaluation was therefore planned to be performed in conjunction with the trainings. As outlined in the Evaluation Framework and Guidelines documents (D8.1a and D8.2a) the major aim and evaluation focus of the first round of evaluations (formative evaluation) was to investigate the usability and perception of the functionality provided by the GRAPPLE authoring tools - in order to gather valuable feedback for developers to improve these tools. This actually constitutes a precondition for the successful authoring of adaptive courses and their application for teaching and learning. In the evaluation guidelines document a procedure of successive training and evaluation stages was suggested (compare Figure 7). This scheme was adopted in the different deployment settings to a different extent and with different focuses, as will be described in the following subsections.
2.2.1 WP9-specific Adjustments

In WP9 it was decided to focus purely on the evaluation of the authoring tools. The learning perspective was consciously skipped; instead an even more detailed evaluation of the authoring tools was chosen to be realised. Consequently, the evaluations carried out were confined to the stages 1 and 3 of the suggested scheme (compare Figure 8).

The very first evaluation of the GRAPPLE authoring tools was realised in terms of a demonstration evaluation aiming at gathering initial and general level feedback on usability and acceptance. Potential users were presented with the authoring tools and asked to give feedback (by filling out a short questionnaire) after a demonstration and trying them out.

The second round of evaluations was carried out in terms of a more in-depth investigation of the authoring tools, i.e. realising a training provider evaluation with interested authors after more intensively examining GAT. These evaluations covered the first and the third stage of the in-depth evaluation: while the first stage addressed the collection of background information about the user (in the scope of a training session introducing the objectives of GRAPPLE and adaptivity in general), the third stage aimed at collecting detailed feedback on functionality, usability, and acceptance. In a first instance this was realised by addressing computer science students taking a course on adaptive hypermedia. These participants intensively worked with the authoring tools over one term and subsequently gave feedback.

This was followed by a set of training and evaluation workshops at different university partner institutions also covering the respective two evaluation stages. In addition to the general aim of investigating usability and acceptance of GAT, however, for these workshops the objective was extended. The perception of and expectation towards the authoring approach used in the conceptual adaptation model (CAM) tool was additionally and explicitly examined. While the creation of a domain model (DM) with the DM tool – due to its similarity to other concept mapping methods (e.g. Novak, 1998; Steiner, Heller & Albert, 2006) – was identified to be rather well-known to potential authors, the CAM approach of structuring and creating adaptive courses is rather uncommon for educational practitioners. The aim of the GAT evaluation therefore was to go beyond the general evaluation objectives (usability and acceptance) and to investigate the intention of authors when modelling a particular CAM visually and the visual expressions behind educational intentions. To tackle this objective, the evaluation methodology was extended by the use of a variant of the think-aloud method (i.e. asking the participants to explain the meaning of their CAMs including the visual arrangement of the model) and subsequent protocol analysis (for more information on this technique please refer to section 4.3.1 The Thinking-Aloud Method and Appendix A1).
2.2.2 WP10-specific Adjustments

In the corporate settings a different approach than in WP9 was taken. In the training and evaluation events carried out in WP10 all three stages of the in-depth evaluation scheme were covered — without, however, investigating in detail the CAM authoring approach as in WP9 (see Figure 9). Originally planned as separate events, in the actual planning of training and evaluation in WP10 it turned out to be more realistic and feasible to combine the different stages into one single event — mainly because of the difficulty to recruit business people and requiring their time for several successive events, but also due to some delay in the availability of stable and ready-for-evaluation software components.

Consequently, a workshop outline was developed covering all three stages of in-depth training and evaluation within one and the same event. Such an event addressed training providers and authors and therefore mainly attempted to gather evaluation data from this perspective (Stage 3). In addition, by making the participants also familiar with the learning perspective of GRAPPLE, the opportunity for a first learner evaluation could be taken (Stage 2). An according training and evaluation event was consequently made up of three stages of training with subsequent evaluations. Stage 1 mainly served a general introduction into the GRAPPLE project and its main objectives, concepts etc. and the collection of demographic and background data from the participants. Stage 2 was on learning with GRAPPLE — the GRAPPLE Adaptive Learning Environment (GALE) was presented and participants could simulate and experience a short learning session by browsing through an exemplary adaptive course. The second part of evaluation constituted therefore the evaluation of the GRAPPLE system from the learners’ perspective. Usability and user acceptance were addressed at this point, as well as adaptation quality in terms of subjective learning effectiveness, cognitive effort, and user feedback. Stage 3 focused on the GRAPPLE authoring tools and constituted a presentation and hands-on authoring experience. The following third part of evaluation aimed at evaluating usability, in general, and GAT functionalities, in particular, as well as user acceptance from the authors’ perspective.
2.2.3 Adjustments for the Visualisation Evaluation

For the evaluation of the visualisation mock-ups developed in T4.5 the basic evaluation methodology as outlined in D8.2a was followed. As outlined in D8.2a, if different visualisations are to be evaluated, the questionnaire actually needs to be filled out for each visualisation.

A few adaptations on the questionnaire presented in D8.2a have been done according to exchange and collaboration with WP4. These updates were especially carried out in order to reduce workload for respondents, as a whole collection of visualisations was to be evaluated.

- The demographic part of the questionnaire was shortened (by omitting questions on profession and experience with learning systems) in order to reduce workload and focus more on the actually relevant part of giving feedback on the visualisations.

- Instead of presenting all visualisations (i.e. for learners and for instructors) to all participants, it was decided to create different paths of the online questionnaire and to consequently query learners only on visualisations intended for learners and instructors only on visualisations intended for instructors. In this way, the number of questionnaires to be filled out by each respondent could be reduced to those referring to visualisations actually relevant for him/her.

- As some visualisations were very similar with only small variations, the survey was shortened by including for a given visualisation a question on whether respondents would find a certain other variant of this visualisation helpful or preferable.

- The questionnaire covering questions on the usability, perceived benefits, and acceptance of each visualisation was adapted. Due to the distinction of the two visualisation types, not all items of the questionnaire were relevant for each visualisation. While the statements on usability and user acceptance remained fully the same for both types, statements on meta-cognition and cognitive load have different variants for learner and instructor visualisations. The items on learning effectiveness and benefits for peers only apply to the learner visualisations while those on benefits for instructors only apply to the instructor visualisations. For details please refer to Table 1.

- The open questions (originally 'what did you like best', 'what did you like least' and 'what should be improved') were reduced to only one question (i.e. ‘Could you please describe in more detail your opinion on this visualisation (positive/negative aspects, how to improve it)?’).
<table>
<thead>
<tr>
<th>Topic/Subscale</th>
<th>Questionnaire for Learner Visualisations</th>
<th>Questionnaire for Instructor Visualisations</th>
</tr>
</thead>
</table>
| Usability (U): | - I find this visualisation suitable for getting an overview of the current status in the learning process.  
- I think the visualisation provides irrelevant information. (R) | - I find this visualisation suitable for getting an overview of the current status in the learning process. (R) 
- I think the visualisation provides irrelevant information. (R) |
| Suitability for the task (ST) | | |
| Usability (U): | - It is easy to understand this visualisation.  
- I find this visualisation unnecessarily complex. (R) | - I think this visualisation can help instructors to reflect on their teaching. 
- I think this visualisation will not significantly promote instructors’ understanding and awareness of the teaching and learning process. (R) |
| Self-descriptiveness (SD) | | |
| Metacognition (M) | - I think this visualisation can help learners to reflect on their learning.  
- I think this visualisation will not significantly promote learners’ understanding and awareness of their learning progress. (R) | - I think this visualisation can help learners to reflect on their learning. 
- I think this visualisation will not significantly promote learners’ understanding and awareness of their learning progress. (R) |
| | | |
| Cognitive Load (CL) | - I think this visualisation is able to leverage mental workload.  
- I think interpreting this visualisation would put additional cognitive effort on the learner. (R) | - I think this visualisation is able to leverage mental workload. 
- I think interpreting this visualisation would put additional cognitive effort on the instructor. (R) |
| | | |
| Learning effectiveness (LE) | - I think the use of this visualisation will not make a difference for learning performance. (R)  
- I think this visualisation can help learners in accomplishing their goals. | - |
| | | |
| Benefits for Peers (BP) | - I think this visualisation would make collaboration among peers more difficult. (R)  
- I think this visualisation can help learners to better understand their learning through comparison with other learners. | - |
| | | |
| Benefits for instructors (BI) | | - I think this visualisation would help instructors in tailoring their teaching to individual learners. 
- I don’t think that this visualisation can help teachers in better understanding their students’ needs. (R) |
| | | |
| Acceptance (A) | - I would make use of this visualisation for my work.  
- I would recommend others to make use of this visualisation. | |
| | | |

Table 1: Adapted questionnaire versions for the different visualization types.  
Note: Items marked with (R) are negatively polled items. For acceptance no negatively polled item was used; for future uses this could be aligned with the other aspects for the reason of consistency.

3 Reflection on the Formative Evaluation Methodology towards Summative Evaluation

Through a steady exchange and cooperation with WP9 and WP10 and provision of support in the preparation, conduction, and post-processing of the evaluation events, needs for adjustments and improvements of the evaluation methodology could be identified. First and foremost, it became clear that the different deployment and evaluation settings feature different conditions for the empirical work with users and therefore pose slightly different requisites with respect to the evaluation methodology. This was translated into according adjustments of the formative methodology to the needs of the respective setting as described in the previous sections. These requisites and contextual conditions of the empirical settings need to be
taken into account also for the summative evaluation methodology and have to be reflected by different approaches or variants tailored to the respective setting.

In the evaluation workshops the participants’ perception of the number and length of questionnaires presented could be experienced. From this it became clear that the empirical evaluations should not put too much workload on the users in terms of filling out questionnaires. Users participated in training and evaluation events voluntarily and were assumed to be on principle willing to provide their feedback. If, however, too many and/or too long questionnaires are presented, the participants’ motivation and commitment to provide reliable and helpful responses will most likely decrease. Although this motivational aspect is not measured explicitly (it would actually mean additional workload), an inappropriately high density of questions/questionnaires might influence answers. This is especially the case when asking for example for providing optional qualitative feedback, which users then tend to abstain from. In the evaluation workshops especially the IsoMetrics (though the short version was used) was perceived to be rather long. Still, the use of this questionnaire was considered valuable due to its multi-faceted consideration of usability and the comparability of results. To reduce the perceived workload in terms of responding for future evaluation events it might be worth considering the exclusion of subscales from the IsoMetrics, which appear not highly relevant for the purpose of evaluating the GRAPPLE tools. To this end, the response rates to the different scales from the evaluations already carried out could be used as a source of information; consistently low response rates on a certain scale can be interpreted in terms of inappropriateness of the respective scale and users therefore being unable to answer the respective questions. Furthermore, upcoming evaluations could be split off into several subsequent events – this would however increase the difficulty of recruiting participants and bear the problem of drop outs. Another option would be to ask participants to fill out (part of) the questionnaires online, outside of a face-to-face setting, and whenever they find time (which would however mean that the conditions of the evaluation procedure cannot be kept completely constant). Furthermore, for future evaluations alternative approaches of data collection should be considered that allow users to give more direct feedback and therefore to gather qualitative data more easily and to a larger extent, like e.g. think-aloud or group discussions (see also section 4.3 Qualitative Evaluation Methods.)

The dedicated focus of the formative evaluation phase was to investigate the usability and acceptance of the GRAPPLE authoring tools, because a successful authoring process is a critical prerequisite for the availability of adaptive courses and the evaluation of adaptive learning experiences. Although already first steps toward learner evaluations have been taken with authors in WP10, it is important for the summative evaluation phase to concentrate especially on the detailed investigation of the benefits of adaptive learning, more concretely the consumption of adaptive courses via the GRAPPLE Adaptive Learning Environment (as stand-alone version or integrated into an LMS). In this regard, not only the usability and acceptance of GALE are of interest, but especially the different aspects of adaptation quality as elaborated in the first version of the evaluation framework (cf. D8.1a, section 4.4) – user feedback, cognitive effort and learning effectiveness. The initial learner evaluations conducted in WP10 basically consisted in a simulation of a learning experience and therefore had to rely on subjective ratings for learning effectiveness. The focus on the learning perspective in the summative evaluation phase will also offer the opportunity of objective learning performance assessment. In this context, also the issue of lifelong learning should be taken into consideration, as the subject of the GRAPPLE project includes lifelong learning (see also section 4.4 Evaluation of Lifelong Learning).

Nevertheless, the evaluation of GAT needs to be continued in summative evaluation, too. While the formative evaluation cycle served the collection of feedback on the initial implementation of the tools and the derivation of ideas and needs for improvement, the summative evaluation should now investigate whether usability, functionality, and acceptance of the tools could be noticeably improved (and to what extent) for the final implementation. One problem in the first evaluation cycle on GAT was that the evaluation participants could not experience how the adaptive course structures they created are actually translated into an adaptive course. This complicated the authoring process as there was no possibility to check whether the educational intentions are appropriately realised in adaptive story lines and authors could not see the actual outcome of their authoring efforts. With the final implementations of the GRAPPLE system and the integration of its different components this can be overcome and thus, training and evaluation events can be further empowered.

In addition, with the evaluation studies conducted on the authoring tools in the first evaluation cycle it may be discussed whether usability could be somewhat confounded with learnability. The evaluations of the tools in the scope of the workshops carried out in WP9 and WP10 took place after short demonstration and training sessions, which might have led to an underestimation of the tools’ usability due to lacking abilities of users on how to use the tools as a result of only short training phases. In case of the training and evaluation that was conducted with computer science students in WP9 the training and concrete experience with the tools was very intensive and took several months. The evaluation results from this study are very similar to the results of rather short time training sessions with subsequent evaluations carried out in WP9 and WP10. It
might therefore be argued that the usability scores of GAT are not expected to considerably change with longer trainings. It needs, however, to be taken into account that the population participating on the longer term training came from a computer science context and do not represent average users. Academics from other fields and representing potential future users of GRAPPLLE might learn to use properly a new piece of software only after a longer period. As a result, longer training periods should be aimed for to ensure having a good basis for evaluation and stimulating the actual use of the GRAPPLLE tools.

In order to ensure efficient training on the GRAPPLLE system, an approach similar to the minimal manual approach (Carroll, Smith-Kerker, Ford, Mazur-Rimet, 1987-1988) should be taken. As far as possible for future evaluations foregoing training phases should last reasonably longer, which could be accomplished by the realisation of distance training phases instead of or complementing face-to-face trainings. The use of online training would increase the flexibility by allowing users to freely choose the time for going through the training material. Moreover, this would bring a further possibility of practical application of GALE and according learner evaluations.

The methodology for the summative evaluation phase also needs to take into account the contextual conditions of the empirical settings by allowing variations and adjustments in the methodology according to the requisites of academia and business. The main difference between the two settings mainly refers to the number of evaluation participants available and the time they are able and willing to spend on evaluation. When addressing learners as users in an evaluation, in a university context it is expectable to have a large number of students available – through recruiting whole courses. In the case of corporate settings it appears more difficult to consult a large number of evaluation participants, especially in case of learners as the target stakeholder group. For the stakeholder group of training providers, which covers people that are responsible for the provision and teaching of courses in a broad sense (i.e. including authors, lecturers, trainers, human resource managers), the number of available evaluation participants will be more equally distributed in both settings. The target user group in this case will be ‘early adopters’, who are more open towards new concepts and are usually more error tolerant than the targeted end-user community. Although this group is able to provide highly valuable information for development processes, the group is commonly the smallest group among the target population.

In addition it has to be taken into account that potential users will only have limited time available for participating in training and evaluation events and have full schedules. This is especially relevant in business contexts. Several successive face-to-face meetings for training and evaluation appear quasi impossible in this case. This also calls for more flexible ways for carrying out trainings and evaluations, e.g. through webinars and online surveys with only limited or no phases of face-to-face attendance. In particular in WP10 the intention has been expressed to apply such online approaches to training and evaluation in order to facilitate the organisation and execution of such events. In any case, future evaluation studies should focus on real teachers and learners representing early adopters from the population of potential future users of GRAPPLLE.

4 Methodological Issues for the Summative Evaluation Phase

On the basis of the first version of the evaluation framework and the experiences made and feedback gained this chapter discusses methodological issues and state of the art in evaluation research that are considered relevant for the summative evaluation methodology of GRAPPLLE.

4.1 Evaluation Topics and Instruments in Adaptive System Evaluation

The evaluation topics addressed in GRAPPLLE evaluation have been elaborated in detail in the first version of the evaluation framework. An overview of possible evaluation instruments has been given and instruments suitable for the investigation of the individual evaluation topics have been discussed. In addition, the state of the art in the evaluation of adaptive systems has been presented. We would like to shortly recapitulate here the most common evaluation topics and instruments used in current evaluation practice of adaptive systems as presented in a literature review by Van Velsen, van der Geest, Klaasen and Steeholder (2008).

User-centred evaluation is considered critical for adaptive systems, as these systems are as such user-centred, aiming in the tailoring of information and interface to the needs of the user. User-centred evaluation is understood as an empirical evaluation obtained by assessing user performance and user attitudes towards a system, by gathering subjective user feedback on effectiveness and satisfaction, quality of work, support and training cost. By this it is possible to verify the quality of a product, to detect problems and to support decisions (De Jong and Schellens, 1997). In consequence this may lead to higher adoption of the system, more ease of use and more pleasant user experience.
In their analysis of evaluation studies on adaptive systems Van Velsen et al. (2008) could identify four main categories of evaluation topics:

1) Variables concerning attitude and experience: appreciation, trust and privacy issues, user experience, and user satisfaction.
2) Variables concerning actual use: usability, user behaviour, and user performance.
3) Variables concerning system adoption: intention to use, perceived usefulness
4) Variables concerning system output: appropriateness of adaptation, comprehensibility, unobtrusiveness.

The most commonly addressed evaluation variable turned out to be usability, followed by appropriateness of adaptation, perceived usefulness and intention to use. The latter evaluation variables actually constitute important aspects of user acceptance. These most frequently addressed evaluation topics actually correspond well to the evaluation topics addressed in GRAPPLE (for a detailed discussion on the evaluation topics in GRAPPLE please refer to D8.1a). The investigation and demonstration of usability is of course crucial in the evaluation of a new technology and therefore also highly important in GRAPPLE. Appropriateness of adaptation refers, as the wording is speaking for itself, to the question whether the adaptation realized by the system is appropriate and corresponds to the users’ needs. This evaluation topic is reflected in GRAPPLE in terms of the concept of adaptation quality. Adaptation quality thereby constitutes a more systematic approach by clearly distinguishing and investigating different aspects playing a role for the appropriateness and quality of adaptation (i.e. learning effectiveness, cognitive load, and user feedback on adaptation quality). User acceptance is also addressed in the evaluation methodology of GRAPPLE. In GRAPPLE also a more comprehensive consideration of user acceptance is taken into account, addressing in addition to perceived usefulness and intention to use also perceived ease of use and attitude towards a new technology as important aspects of user acceptance according to Davis’ (1986) Technology Acceptance Model.

Evaluation designs commonly used in the user-centred evaluation of personalised systems can be summarised as comparative studies and observational studies in laboratory/real-life settings (Van Velsen et al., 2008).

Comparisons between personalized versus non-personalized systems try to identify the difference between a personalized version of a system and one without the personalizing feature. In most cases the user performance with two versions of a system is measured to judge whether the personalized system was better for the user than the non-personalized system. According to Höök (1997, 2000) who discussed the validity and reliability of such comparisons in detail, the comparisons between a personalized system with one where the personalizing feature has been removed is deemed a false comparison because the system without the personalizing feature is no longer a worthy opponent. Comparing personalised systems and non-personalized systems concerning their user performance, the involved non-personalized system should be a worthy equivalent for the personalized system and not a weak version of the original system. In GRAPPLE comparative studies contrasting an adaptive version with a non-adaptive version of the system are also foreseen. Through the integration of the GRAPPLE Adaptive Learning Environment with LMS, however, a fair form of such a comparison can be realised (compare also section 5.5.2.3 of this document and section 5.4.1 System as a Whole in D8.1a)

In laboratory and real-life observations the researcher observes a participant using and working with a personalized system and notes interesting events or records the whole session. Using a laboratory allows the researcher to control the environment by excluding outside influence. Therefore the researcher can focus on the assessed variables. The problem of using a laboratory setting is the reduced ecological validity because of losing the real-life setting. This means that the situation in a laboratory is not identical to a real-life setting, so the results may not be generalisable. In GRAPPLE both, laboratory-like as well as real-life settings are envisaged to be involved in summative evaluations, although no observational studies in the narrower sense are foreseen in both contexts. On the one hand, in evaluation workshops participants will be recruited to take part in training and evaluation in an unfamiliar setting separated from their daily working environment and practice. On the other hand, deployments are planned in which the GRAPPLE system is intended to be used in educational practice (compare also section 5 Summative Evaluation Methodology in GRAPPLE).

Van Velsen et al (2008) present in their literature review also a summary of the most commonly used evaluation methods. These are questionnaires, interviews, log data analysis, focus groups and group discussions, thinking-aloud, and expert reviews. A concise overview of these types of evaluation instruments and other methods as well as their advantages and drawbacks has already been given in D8.1b (section 3.2 Methods and Techniques for evaluation). Here we only want to mention that also the evaluation instruments used or planned to be used in GRAPPLE evaluation are congruent with the current evaluation practice. In
the requirements analysis phase interviews have been conducted. In formative evaluations mainly questionnaires were used, and this instrument is going to be applied as the major evaluation method in summative evaluation, too. In addition, we have identified the need for applying also qualitative methods to a larger extent, in order to gain a deeper understanding of the users’ opinion and their thinking and interaction processes with the system. As a result, focus groups and thinking-aloud were identified as valuable evaluation methods. These methods will be discussed in more detail in section 4.3 Qualitative Evaluation Methods.

Van Velsen et al. (2008) suggest orienting evaluation in general on the iterative design process of an adaptive system – accordingly the goals of evaluation will change from supporting decisions on design, to identifying problems and issues for improvement, to verifying the quality as an adaptive system becomes more and more mature (see Figure 10). The phases suggested by this model can actually be referred to as requirements analysis, formative and summative evaluation (e.g. Gena & Weibelzahl 2007; Harvey, Higgison, & Gunn, 2000). These phases of evaluation aligned with the development process have actually also been realised in GRAPPLE. While requirements analysis and formative evaluation phase have already been concluded, the summative evaluation investigating the benefit of the final implementation of the GRAPPLE system will be done in the final phase of the project.

Figure 10: Design process and associated evaluation goals and phases (adapted from Van Velsen et al., 2008).

4.2 Interpreting Usability Scores

Regarding the scores of questionnaires assessing the usability of a product it is often difficult to interpret the data in an appropriate way. Often the results of these questionnaires are not reported systematically. Furthermore, the information describing how the numeric scores translates into an absolute judgment of usability is not known. Often, questionnaires do not provide reference values and especially in case of newly created questionnaires no clear score cut-offs are defined. For this reason it is difficult for the reader to judge the quality of an evaluation. In a well-conducted evaluation study optimally the validity and reliability of the instruments used should be reported (e.g. The Institute of Higher Education Policy, 1999). In particular, reporting on the means and the standard deviations for each variable is important when describing evaluation results and can provide the reader with a clear view on the answers the participants gave and their consensus on a certain topic.

Aside from a proper reporting of results on the usability questionnaires applied in GRAPPLE evaluation, we want to concisely discuss here the issue of interpreting usability scores for the two main usability questionnaires used in GRAPPLE.

4.2.1 System Usability Scale

In GRAPPLE evaluation the System Usability Scale (SUS; Brooke, 1996) is used for gathering a quick general subjective assessment of usability (in demonstration evaluations – compare 5.3 Demonstration Evaluations). The SUS consists of ten items that have to be answered on a five-point rating scale. The answers on these ten items are then used to calculate an overall usability score. This score has a range from 0 to 100.
While a 100-point scale is intuitive in many respects and allows for relative judgments, information describing how the numeric scores translate into an absolute judgment of usability is not known. In an attempt to investigate and determine what individual SUS scores mean, Bangor, Kortum and Miller (2009) therefore tried to solve this problem by adding a seven-point adjective-anchored Likert scale as an eleventh question. This Likert scale is used to determine if a word or phrase could be associated with a small range of SUS scores. Figure 11 shows the adjective rating scale.

![Adjective rating scale added to the SUS](adopted from Bangor, Kortum and Miller, 2009, p.117).

To determine how well the ratings using the adjective rating scale matched the corresponding SUS scores given by the participants a correlation analysis was conducted. The results show an extremely well correlation between the Likert scale scores and the SUS scores with $r=0.822$ ($\alpha<0.01$). The mean score for each adjective rating is shown in Figure 12.

![Adjective Ratings vs. SUS Scores](adopted from Bangor, Kortum and Miller, 2009, p119).

According to Bangor et al. (2009) the finding that the adjective rating scale very closely matches the SUS scale suggests that it is a useful tool in helping to provide a subjective label for an individual study’s mean score. This means that the addition of the adjective rating scale to the SUS may help practitioners interpret individual SUS scores and aid in explaining the results to non-human factors professionals.

### 4.2.2 IsoMetrics

The IsoMetrics questionnaire corresponds to a user-oriented approach in software evaluation on the basis of ISO 9241 (part 10). It comprises 75 items assembled in seven subscales corresponding to the seven design principles of ISO 9241 (Part 10):

- Suitability for the task
- Self descriptiveness
- Controllability
- Conformity with user expectations
- Error tolerance
- Suitability for individualisation
- Suitability for learning.
With the aid of the IsoMetrics it is possible to gather and use information within an iterative software development cycle. Two versions of the questionnaire assist – one gathering additional qualitative data for the identification of malfunctions in the context of formative evaluation and one gathering scores on the different usability dimensions only.

The IsoMetrics questionnaire has been proven to be a reliable and valid tool. Furthermore it has shown its practicability in several software development projects. IsoMetrics gives efficient support to identify weak points of software systems, and therefore provides concrete impact on the improvement and redesign of software systems (Gediga, Hamborg & Dünstsch, 1999). Furthermore, for the subscales comparison values are reported, i.e. values which have been assessed for popular application software (e.g. LaTeX, Word, or SAP/Term) (see Technical Report “Normtabelle zum Isometrics-Verfahren” from Gediga, Hamborg, & Willumeit (1996) on http://www.isometrics.uni-osnabrueck.de/papers.htm). The majority of these values are around 3.5. Based on this, cut-offs in terms of a criterion-based evaluation for the GRAPPLE tools have been set according to the following rationale:

- Values of 3.5 and higher can be seen as a good to very good evaluation in the sense that the tools can compare with popular software.
- Values between 2.5 (middle of the scale) and 3.5 are suggested as a medium to good evaluation.
- Values below 2.5 indicate a rather poor evaluation and suggest that several improvements have to be made.

To increase the comparability and vividness of the evaluation results gathered with IsoMetrics it is also possible to create a comprehensible illustration of the different categories assessed by the IsoMetrics in form of a profile (see Figure 13 and Figure 14 for examples). Such a usability profile gives an immediate idea on stronger and weaker aspects of a technology. When visualizing the profiles of different software or different software versions, pros and cons of the different alternatives become apparent. When considering the example depicted in Figure 14, for instance, it can easily be seen that both software versions basically feature a similar usability profile, but software B is characterized by weaker results in self descriptiveness and controllability.

![Figure 13: Example usability profile of IsoMetrics results.](image-url)
4.3 Qualitative Evaluation Methods

There is much debate in the field of evaluating software and learning technology about the appropriate balance between quantitative and qualitative methods. The general tenor of researchers is to use a mixed or hybrid model, i.e. a combination of qualitative and quantitative techniques appropriate for the specific evaluation questions to be examined (Harvey, Higgison, & Gunn, 2000; Mandinach, 2005). In general both qualitative as well as quantitative approaches are appropriate, but for different purposes. Quantitative methods are able to give precise measurements of outcomes. On the contrary, qualitative methods tend to be less precise, but they can provide rich information on the details of processes, e.g. the interaction and work with a piece of software. Consequently, qualitative methods are suitable for gaining an understanding of these processes.

In the formative evaluation phase primarily questionnaires have been applied for gathering evaluation data. Standardised questionnaires allowed general insights on the satisfaction and expectations of people towards the GRAPPLE tools. Standardised and closed-ended questionnaires were complemented by newly created questionnaires and especially open-ended questions aiming at gathering qualitative data. Though a considerable amount of valuable qualitative feedback could be gathered with the open questions, this type of question is in general rather laborious to answer. The requested responses in terms of open feedback of course require longer explanations than simply providing an answer on a rating scale. Providing extensive written explanations on opinion and experiences therefore might be experienced somewhat uncomfortable by evaluation participants. Therefore, more direct ways of expressing this kind of feedback appears preferable. The use of qualitative methods in this case would be able to gather even more detailed feedback and allow a responsive data collection and therefore clarification or going into detail on particular issues.

As the information provided by questionnaires offers only limited support for guiding the development of an actual system and it does not allow gathering specific requirements or intentions, the formative evaluation methodology has been complemented in WP9 by a variant of the think-aloud technique (cf. section 2.2.1). User requirements and intentions are often tightly interwoven with the practice in which a tool is applied. To capture more detailed information of this practice more qualitative evaluation approaches are required. This information can identify conceptual problems in the design of a tool; provide domain and task specific requirements; and surface intentions and applications that go beyond the original intentions of the development process. Such information helps to make design decisions for the improvement of tools with regard to the actual needs of the target audience. Moreover, it allows the target audience to influence the development process.

The summative evaluation in GRAPPLE aims at investigating and demonstrating the final achievement of GRAPPLE and benefit of the creation and personalised presentation and consumption of adaptive learning content. As such, this evaluation objective argues for quantitative methods. It is desirable to get to know about the outcomes of using the GRAPPLE system, but nevertheless also to get an understanding about the processes of interaction while authoring and learning. For the latter aspect the application of qualitative methods is of interest. Qualitative methods shall in this case not specifically aim at gathering information for guiding further development, but rather at understanding how people are using and interacting with the GRAPPLE tools and triangulating this with quantitative evaluation data.
A variety of different qualitative methods and different variants of those methods exist, e.g. introspection, observation, thinking-aloud, interviews, and focus groups. These methods basically differ in the experience and expertise required to carry them out, but also with respect to data analysis. In the following sections two qualitative methods shall be discussed in more detail, which have been identified as especially suitable for the evaluation in GRAPPLE.

4.3.1 The Thinking-Aloud Method

According to Someren, Barnard, and Sandberg (1994) the think-aloud method is characterised by asking people to think aloud, i.e. to express their thoughts, while solving a problem. While observing the users, they are encouraged to articulate whatever they are thinking, doing, and feeling while performing a task in question. By doing this, the user just has to concentrate on the task: he/she is asked to give a concurrent account of his thoughts without any interpretation or explanation of his/her actions. This talking while solving a problem is performed nearly automatically and does not interfere with the task performance (Ericsson & Simon, 1993). Observers take notes of everything that the user says, without interrupting him or interpreting or justifying his actions. Alternatively or in addition, the test session can be audio- and/or video-recorded. This allows the observer to complement the observation notes or to go back e.g. to see how the user reacted in a specific situation.

The method of thinking-aloud allows to find the specific problems a person has by interacting with an interface by delivering a high-quality, qualitative user feedback. The data gathered in this way allow understanding the cognitive processes of the user by making explicit what is implicitly present in users who perform a specific task.

According to Hannu and Pallab (2000) two different experimental procedures on the think-aloud method can be distinguished: the concurrent thinking aloud protocol and the retrospective thinking aloud protocol. The first one is collected during the execution of the task. In contrast, with the second procedure think-aloud data is gathered after having accomplished a task. For further information on the thinking aloud method, its use in software and e-learning evaluation, strengths and weaknesses, as well as the data analysis of think aloud please refer to Annex A1.

4.3.1.1 The Thinking-Aloud Method in the Context of GRAPPLE

In the context of GRAPPLE the think-aloud technique is applicable for both, evaluations from an authoring as well as from a learning perspective.

The first case, i.e. the use of think-aloud for GAT evaluation, has already been realised in the formative evaluation phase of WP9. With regard to the original approach of think-aloud verbalisations the situation in GRAPPLE is slightly more complex. For authors GAT is the front-end to the GRAPPLE infrastructure that allows them to express their adaptive instructional designs. Therefore, the focus of the usability evaluation has to go beyond the evaluation of the interfaces and has to include the educational rationale of the results into the analysis. In other words, the formative evaluation GAT not only covers whether the participants can familiarise with the authoring environment, but it also includes the question regarding the expected educational processes and adaptive behaviour for a given instructional design. Because of this additional objective for the evaluation, the original think-aloud method has been altered: instead of capturing only the usability problems with the authoring environment, the evaluation procedure focuses on the instructional intentions that were explicitly (as XML) or implicitly (only available graphic presentation) modelled. In order to enrich the verbalised information, the graphical arrangements of the instructional designs are also collected. This allows identifying patterns of similar graphical notations for instructional designs in order to provide end-user requirements for the improvement of the GRAPPLE infrastructure. When applying the original method of thinking-aloud during the actual use of GAT, information on the involved working and thinking processes and probably occurring issues during authoring can be gathered.

In the second case, i.e. applying thinking-aloud for learner evaluations, a deeper understanding and insight on the adaptive learning processes while using GALE can be gained, for example information on the learners' perceived relevance of adaptations. The think-aloud method may be used to explore in detail the ways in which students engage with the learning environment and the learning resources presented, and to evaluate the ways and extent to which this might enhance their learning.

4.3.2 Focus Groups

There is a wide range of definitions of a focus group in the literature. In social research characteristics like organised discussion, collective activity, social events and interaction are significant contributions given by
focus groups. According to Powell and Single (1996) a focus group is a group of individuals that is selected
and assembled by evaluators or researchers in order to discuss and comment on the topic of interest on the
basis of their personal experience. Similarly to this, Morgan (1996) defines a focus group as a research
technique that gathers data through group interaction on a topic determined by the researcher. This
definition is based on three basic components. First of all, focus groups are a research method to gather
data. Second, the interaction in a group discussion is the source of these data. This means that people are
couraged to talk to one another not only to respond to a question in turn asked by the researcher. Third,
the researcher plays an active role in creating the group discussion for data collection purposes.

The main idea behind the focus group method is that group processes can help people to explore and clarify
their views in ways that would be less easily accessible in an interview. Well worked group dynamics can
take the research in new and often unexpected directions by bringing out users’ spontaneous reactions and
ideas.

In interactive system development and evaluation focus groups are a powerful tool to discover what users
want from the system and how they operate the system. For further information on the focus group method
please refer to Annex B1.

4.3.2.1 The Focus Group Method in the Context of GRAPPLE

The focus group technique has implicitly already been applied in the scope of the training and evaluation
workshops. By having a small special interest group of authors taking part at the event, naturally group
discussions emerged in the course of the workshop, while being presented with the GRAPPLE tools and
trying out their functionality. This can be considered as a ‘less structured’ discussion than an actual focus
group. The trainer and evaluator in general did not strongly control the discussion, but rather tried to slightly
guide it; the issues discussed and the extent of discussion was more or less open to the participants. As the
focus group approach was not explicitly part of the evaluation methodology, no systematic data recording
and analysis on the respective discussions was realised in terms of audio-recording, transcribing, and coding
data. Rather, the discussion contents were recorded by taking notes and subsequently summarising the
most prominent issues. Therefore, they served as an additional, informal information and feedback source
on GAT.

For summative evaluations the focus group method should be taken up explicitly into the evaluation
procedure in order to ensure a systematic recording of group discussions. Focus groups are considered to
be a very suitable instrument for gathering feedback on the functionality, opinion, and acceptance of the
GRATTLE system and allow a responsive procedure by asking for further explanation and reasons. A
disadvantage is the complexity of data analysis and the fact that the opinion of each individual can hardly be
identified and thus, cannot be related to other evaluation variables (e.g. learning effectiveness). In any case,
the application of this evaluation technique is straightforward in case of the conduction of evaluation
workshops similar as in the first evaluation phase, as these events by nature provide the appropriate setting.
For a more efficient data collection the objectives of the group discussion should be clarified and specified in
the evaluation methodology, in order to allow a more consistent guiding and moderating of the focus group
across different workshops and partner institutions. Furthermore, the data recording technique needs to be
explicitly specified and required in the evaluation guidelines, which will most probably consist in taking notes
and additional, optional audio-recording. Based on this, data analysis can be carried out and the contents of
the group discussion can explicitly feed into the evaluation report. In this way, the outcomes of the focus
groups may contribute to a better understanding of users’ perception and opinion of the GRAPPLE system
and the reasons behind.

When thinking about an increased use of online training attempts allowing more flexibility in the participation
on training sessions, in particular one further variant of the focus group method appears highly relevant for
GRAPPLE purposes: online focus groups. This is basically the conduction of a focus group in an online
setting (Gaiser, 2008). It allows the relatively easy and inexpensive conduction of group discussions, with the
possibility of participants taking part from different places and even at different points in time. Such an online
focus group may be realised in different ways and with different technologies; synchronous or asynchronous
means of communication are applicable – e.g. by the use of chat or online conferencing tools vs. the use of
email or online forums for the group discussions. Online focus groups have proven to be a viable alternative
to face-to-face focus groups (Reid & Reid, 2005). The realisation of an online focus group by the use of
asynchronous communication techniques would provide a viable way of gathering qualitative feedback on
GRAPPLE tools in a flexible and responsive way.
4.4 Evaluation of Lifelong Learning

In the literature a wide range of different evaluation methods for investigating learning effectiveness is discussed. Unfortunately, these conventional methods most often only delve into the aspect of short term learning, while the aspect of long-term learning is neglected, not to speak about the perspective of lifelong learning. In the GRAPPLE project lifelong learning plays an important role and is a dedicated objective of the project. In a project of limited duration, of course, the proper evaluation of lifelong learning is in principle impossible. Evaluating how the GRAPPLE environment contributes to lifelong learning would need to be the subject of another project. Nevertheless, in the second year project review the reviewers encouraged the consortium to consider the aspect of lifelong learning in evaluation. For this reason, the following section gives an overview on the topic of evaluating lifelong learning. Subsequently, the steps that can be taken in GRAPPLE as a timely limited project to incorporate the issue of lifelong learning evaluation are outlined.

4.4.1 State of the Art in Evaluating Life-long Learning

Evaluation plays a central role in any educational system; consequently, procedures and approaches that aim at promoting lifelong learning actually require evaluation attempts to investigate and focus on this lifelong learning aspect, too. On the other hand, however, as evaluation can only be done in a limited scope it will usually not be possible to conduct a comprehensive evaluation in that direction, particularly in case of programmes or projects of a limited time frame.

The evaluation of the benefit of an educational programme or technology in terms of precisely determining long-term effects is very difficult. As lifelong learning is not directly tied to an institutional context it requires a very large and holistic perspective, also for evaluation (Tuinman & Boström, 2002). The concept of lifelong learning can generally be seen as of a very comprehensive and all-embracing nature. This, however, leads to the fact that lifelong learning evades clear definition and constraints and is therefore not directly amenable to evaluation. When attempting to evaluate lifelong learning, this can therefore only be done when breaking it down to a more concrete level of specific educational approach, programme, or technology. Knapper and Cropley (2000) emphasise the importance of monitoring educational developments on a continuing and formative basis. In this way aspects that are successful in promoting lifelong learning can be revealed as well as factors that may mitigate against it may be identified. Therefore, the assessment of life-long learning needs to be done not only in terms of a retrospective, summative judgement on whether a certain approach worked or was better, but should also be done on an ongoing basis to provide information that is rich enough to improve the programme or technology (Knapper & Cropley, 2000).

In order to identify possible approaches of tackling with the issue of evaluating lifelong learning in GRAPPLE, other European projects that have been or are addressing a lifelong learning objective have been looked at in order to identify possible lifelong learning evaluation frameworks. In the following, an overview of related work in two EU-projects, TENCompetence and the current LTfLL (Language Technologies for Lifelong Learning) project will be given.

In the context of the project TENCompetence a technical and organizational infrastructure for lifelong competence development was designed by integrating different models and tools for competence development into a common framework (Schoonenboom et al, 2007). The project covered research identifying a set of problems in lifelong learning, hypothesising that a specific set of functionalities can improve the provision of lifelong learning, developing an infrastructure to deliver those functionalities, and evaluating different pilots which represent the variety of contexts in which lifelong competence development takes place to validate this infrastructure and hypothesis. The evaluation of these different pilots was very complex due to large differences among the pilot sites with regard to the different learning objectives and intended audiences and thus, with different types of learners involved. For this reason a pilot evaluation model was developed for underpinning the evaluation design to get and hold a firm grip on such a complex evaluation. In Figure 15 the pilot evaluation model, showing the implementation steps and evaluation of the learning outcomes is presented (Schoonenboom et al, 2007). This model is understood as a form of implementation evaluation, in which the chain of steps between vision and implementation is described. The model is subdivided into three pillars: a series of implementation steps, a distinction between envisaged and actual mechanisms and outcomes, and a recognition of various sources of differences. In this way a description of the pilot at different levels is enabled, such as “how was the infrastructure implemented”, “how was it used”, “what are the results of use”.

The model builds the basis of a form of implementation evaluation (Patton, 1997), in which the chain of steps between vision and implementation is described (King, Morris, Fitz-Gibbon, 1987; Williams & Elmore, 1976). It helps to clarify, and therefore to handle differences between different stages of implementation and differences between pilot sites. In most cases program evaluations more or less use an impoverished model
which only consists of the envisaged infrastructure on the one hand and the learning outcomes on the other hand. Such a model mixes up differences between several steps of implementation and so the error increases. In a multisite situation it is important to use an elaborated model to pinpoint the differences and thereby make them manageable. For a comparative analysis it is necessary not only to state that the implementation at different sites will be different but also to define them precisely. In summary, this evaluation model focuses on differences which include the discrepancies that arise when one moves down the implementation steps represented in Figure 15. The model also plays an important role focusing the program theory and in testing this theory. Parts of the model are on the one hand the envisaged use of the infrastructure and the envisaged outcomes and on the other hand the envisaged mechanisms that relate the use of the infrastructure to the outcomes. The corresponding evaluation which validates this program theory is called theory-driven evaluation (Chen & Rossi, 1987; Patton, 1997). Another major role of the model is in focusing the evaluation. As the implementation of the elements of the model can be done in so many ways, the elaboration of the model leads to a large list of opportunities for evaluation. Having this list is a first step in focusing, which consists in selecting the most important possibilities from the list (Schoonenboom et al., 2007). A consolidated view of all this factors indicates that this theory-driven model for multi-site pilot evaluation presented in the context of the TENCompetence project can be useful in evaluating a technical and organizational infrastructure for lifelong competence development.

![Figure 15: Pilot evaluation model of TENCompetence, showing the implementation steps and evaluation of learning outcomes (adopted from Schoonenboom et al., 2007, p.44)](image)

In the context of the current LTfLL (Language Technologies for Lifelong Learning) project next-generation support and advice services will be created to enhance individual and collaborative building of competences and knowledge creation in educational and organizational settings. The project makes extensive use of language technologies and cognitive models in the services. The research activities include a validation structure that shall ensure rigorous evaluation in realistic settings. The project is still ongoing and so the criteria for evaluation as described in the project deliverable report (Armitt et al., 2008), will emerge with the development of the scenarios. The evaluation criteria are expected to include:

- Pedagogic/andragogic effectiveness and need, including the extent to which the services foster learning
- Usability, user acceptance and efficiency
- Transferability between different systems and organizational/educational environments
The evaluation topic relevant in the context of evaluating lifelong learning is the first one. An important aim of the LTfLL project is to validate the pedagogic/andragogic effectiveness and need of the services. In the context of the LTfLL project pedagogy is understood to mean instructor-led learning (i.e. teaching) and andragogy to mean facilitated self-directed and/or experiential learning. The purpose is to develop tools for lifelong learning, which requires an understanding of adult learning (andragogy). The indicators of pedagogic/andragogic effectiveness should include three main topics. First the accuracy of the output of the LTfLL service compared with that of the domain experts, second the perception of domain experts and students concerning whether the services meet their pedagogic/andragogic requirements, and last but not least the effectiveness of learning. The objective of validating and evaluating pedagogic/andragogic effectiveness is to establish the extent to which the services address stakeholder values or problems with regard to learning and teaching.

### 4.4.2 Life-long Learning Evaluation in the Context of GRAPPLE

Among the objectives of the GRAPPLE project is the support of life-long learning, through the provision of an adaptive, technology-enhanced learning environment that is usable and guides the user throughout a life-long learning experience. The aspect of lifelong learning should therefore also be reflected in the evaluation methodology of GRAPPLE. Although of course the actual evaluation of life-long learning in GRAPPLE is not possible and would need to go far beyond the scope of the project, it should be taken into account and addressed how the facilities provided by GRAPPLE can support and actually contribute to lifelong learning.

As can be seen from the previous subsection the evaluation of lifelong learning is very challenging – lifelong learning in general is an all-embracing concept, lifelong learning refers to learning over long durations, and learning settings applying educational technology aiming at supporting lifelong learning may be very diverse. These characteristics complicate evaluation and make it a very difficult task. The evaluation model as developed in the TENCompetence project, though caring for the differences in different pilot and evaluation settings, does not directly focus the lifelong learning aspect. In addition, this evaluation model is characterised by accompanying the different implementation steps and therefore can be regarded rather as a framework covering especially also formative evaluation. Consequently, a similar evaluation approach appears not highly relevant and directly adaptable to the summative evaluation phase of GRAPPLE. The evaluation criteria targeted in the LTfLL project in sum are quite similar to those selected and addressed in GRAPPLE, i.e. targeting usability and acceptance, on the one hand, and pedagogical effectiveness, on the other hand. In the current status of this project, however, there are no concrete indications on how learning effectiveness will be evaluated and whether this evaluation will sufficiently reflect the lifelong learning aspect.

In the summative evaluation phase of GRAPPLE the aspect of lifelong learning shall be taken into account from different viewpoints. On the one hand, users’ opinion on the potential of the GRAPPLE system to support lifelong learning can be evaluated. In contrast to approaches that try to find out the learning power of learners and how learners perceive themselves in relation to different dimensions of lifelong learning (e.g. Crick, Broadfoot, & Claxton, 2004), in this case the potential of an educational technology to support lifelong learning will be targeted. This lifelong learning potential can be regarded as an additional aspect of the evaluation topic learning effectiveness. On the other hand, steps toward lifelong learning evaluation shall be taken in terms of actually investigating learning effectiveness. Clearly, a comprehensive evaluation in this direction would need to be the subject of another project. Still, steps can be taken in terms of investigating more long-term learning effects instead of performing learning assessment only directly after a learning phase. This is envisaged to be realised in the context of the system deployments. Especially in university settings an evaluation of longer term learning effects appears conceivable, as lecturers interested in the practical use of the GRAPPLE system will be involved in training and evaluation. It is planned to recruit a small special interest group of instructors stemming from different partner institutions, who are willing to realise teaching of parts of their courses by the use of GRAPPLE. This would mean that a part of the learning content of a whole course would be implemented and taught in form of an adaptive course. This course would be implemented by the use of GAT, allowing the training on the authoring tools and the collection of feedback from the involved instructors. By actually using the adaptive courses as a complement to the rest of the learning content provided in a lecture, a learning assessment at the end of the overall course could be done. This would allow the investigation of learning performance over a longer period of time (possibly in addition to a short time learning assessment) and also allow a comparison to traditionally taught learning contents. In this way, the aspect of lifelong learning could be included to some extent in the GRAPPLE evaluation process.
5 Summative Evaluation Methodology in GRAPPLE

As already mentioned in section 4.1 Evaluation Topics and Instruments in Adaptive System Evaluation, evaluation should be aligned with the planning, development, and implementation aspects of an e-learning technology. Lam and McNaught (2008), for example, support this idea and elaborate the multiple roles evaluation can have in a three-layered cyclic model of e-learning development and evaluation. This cycle model describes the multiple contributions that evaluation can make, concentrating on the contributions of summative evaluation. Summative evaluation makes visible the strengths and weaknesses of a finished e-learning technology. Moreover, summative evaluation can lead to new ideas on more effective e-learning strategies. Lam and McNaught emphasize the potential of summative evaluation to act as a basis of the refinement of the evaluation strategies used in future studies. Accumulation of evaluation experiences and reflection of these experiences can contribute to the advancement of evaluation methodologies, and the tools and strategies used. In GRAPPLE, summative evaluation shall mainly investigate and demonstrate the benefits and achievements of the final implementation of the GRAPPLE components and system. Through reporting on the evaluation outcomes, GRAPPLE’s summative evaluation can furthermore also stimulate and feed into future research and development on adaptive e-learning systems.

5.1 Objectives

As mentioned already in the course of the foregoing sections, the main objective of the summative evaluation phase in GRAPPLE is to demonstrate the benefit of creating and consuming adaptive courses by the use of the GRAPPLE environment. Hereby, the final evaluation cycle shall pay special attention to the evaluation of learning effectiveness and adaptation quality. The main target audience from this viewpoint are therefore learners with different backgrounds and from different educational settings. In the centre of evaluation stands the GRAPPLE Adaptive Learning Environment (GALE), i.e. the environment where the adaptive learning experiences come about. Apart from learning effectiveness and adaptation quality, also the evaluation of learners’ acceptance and subjective usability assessments are of interest.

In addition, the evaluation of the authoring tools is to be continued, as they constitute the part of the GRAPPLE system that is the crucial prerequisite for the realisation of adaptive courses. General usability and acceptance measures shall be collected again to allow a comparison between initial and final implementations. The major evaluation effort will therefore be on the core authoring components. The specific authoring tools for virtual reality and simulations are not part of the core components and will therefore not be in the focus of GAT evaluation.

The GRAPPLE visualisation tools should be targeted in order to evaluate their visualisation quality and expected benefits for users, but also for subjective assessments of acceptance and usability. This evaluation is envisaged to be carried out in the context of learner evaluations with GALE, where the visualisations are used in their intended context and application.

5.2 General Evaluation Scheme

Similar to the first evaluation cycle, summative evaluations in the empirical settings of WP9 and WP10 are planned to be realised in conjunction with according training events. To ensure the efficient gathering of informative evaluation data, it is a necessary prerequisite that users are familiar with the functionality and handling of the GRAPPLE tools and services. Evaluation sessions will therefore be carried out in alignment with and subsequent to training activities and will target the same participants.

In formative evaluation two different levels of evaluation have been realised, general level evaluation and in-depth evaluation. This scheme will on principle be carried on also for summative evaluation. General-level evaluation especially relates to demonstration evaluations in the context of probable conference demonstrations and tutorials on the GRAPPLE tools. In-depth evaluation includes all specially organised and realised GRAPPLE evaluation events aiming at a more intensive contact with prospective users and gathering their feedback after concrete experiences in using and working with the tools. Thereby, a two-tier system of evaluations (and accompanying trainings) is planned. On the one hand, workshops similar to those conducted in formative evaluation are suggested to be realised at different partner sites. These workshops will include the training on the functionality and use of the GRAPPLE components in question including hands-on experiences and a subsequent evaluation session. On the other hand, deployments of the GRAPPLE system are planned. These deployments refer to concrete applications of GRAPPLE in educational practice and shall realise a longer term interaction covering again training and subsequent evaluation phases. Figure 16 provides an overview of the scheme proposed for the summative evaluation phase of the GRAPPLE project. The subsequent sections will outline in more detail the aims and procedure.
of each of the depicted types of evaluation and will elaborate further on the different tiers of in-depth evaluation.

Figure 16: Summative evaluation scheme for GRAPPLE.

5.3 Demonstration Evaluations

5.3.1 Objective
The general objective and methodology of demonstration evaluations representing the general level evaluation part are the same as outlined in the first version of the evaluation framework and guidelines: the collection of general level feedback on usability and acceptance of the GRAPPLE tools and system.

Demonstrations or tutorials on the GRAPPLE system in the scope of conferences and workshops provide an additional opportunity to collect user feedback for the purpose of evaluation. The data gathered can only provide a very rough idea on users’ perception of GRAPPLE’s facilities, but will nevertheless provide useful information on the users’ impression of the tools they got during the demonstration.

5.3.2 Method
The target audience of conference demonstrations or tutorials can be assumed to cover both stakeholder groups, learners and training providers. Correspondingly, potential future users in general (authors as well as learners) are targeted in a demonstration evaluation.
Demonstrations will usually be on certain components or toolsets of the GRAPPLE system, i.e. the authoring tools, the learning environment, or even the visualisations, but may also be on the GRAPPLE environment as a whole. Thereby, the demo will likely be restricted to a demonstration only without direct interaction between system and user, but may also involve the opportunity for users to try out the tools by themselves (e.g. in a tutorial). In any case, the evaluation will address those components of the GRAPPLE systems as presented in the respective demonstration session.

The extent of evaluation data that can be gathered in this scope is of course limited and also the time available for doing the evaluation as such (due to time constraints of the demo session itself). This needs to be taken into account when planning the data collection (e.g. choice of measurement instrument). An evaluation instrument that can be applied in this scope needs to be simple and easy to use; a questionnaire is therefore the best and basically only option. The general idea is to reserve a small amount of time at the end of the demo or tutorial session for the evaluation and to ask participants to fill out the questionnaire. The questionnaire needs to be rather short, containing only a small set of questions or items that are easy to be answered. Due to the limited timeframe available preferably a questionnaire with closed-ended questions (at least most of them) should be applied.

For demonstration evaluations carried out in the final project phase therefore the System Usability Scale (SUS; Brooke, 1996) is suggested to be used – as it has been done in demo evaluations conducted so far. It is a simple, ten-item scale providing a global view of subjective assessments of usability. Because of its brevity and the easy answering mode through a rating scale the questionnaire is easy to administer. To collect also data relating to user acceptance aspects, the SUS has been complemented for GRAPPLE purposes by two additional items of the same type and answering mode like the SUS items. In case of a whole-day tutorial it is conceivable to add also few open questions (e.g. things liked best/least), as in this case most likely a little bit more time will be available for evaluation.

The realisation of the same basic evaluation procedure using the same evaluation instrument will allow comparing evaluation results for the final implementation with the outcomes achieved for the initial implementations of the GRAPPLE components.

5.4 Authoring Perspective: Evaluation of GRAPPLE Authoring Tools

5.4.1 Objective

The main objective of continuing the evaluation of GAT in summative evaluation is to investigate differences between the initial and the final implementation of the tools. Formative evaluation served the collection of user feedback and recommendations as a basis for deriving ideas on the improvement of the tools for their final implementation. These improvements of the tools should actually lead to an improvement of users’ experiences and thus to improved results regarding usability and user acceptance.

Evaluations of GAT shall in particular address those areas that were considered weak during the formative evaluations. Therefore, the following areas are considered important to be considered: usability and simplicity of GAT, pedagogical semantics of GAT, user model variable editing, deployment to GALE. Consequently, changes in usability and acceptance scores gathered by questionnaires are of interest. Further aspects of interest as listed above may be investigated through the use of qualitative methods. Qualitative data, though, is intended to be gathered only to a limited extent, as after summative evaluation there will not be an opportunity for further improvement of the tools within the project lifetime.

Because of the intended comparison between initial and final implementation and the importance of these tools the main objective of GAT evaluations will be to investigate the quality of the core authoring tools, particularly the DM and the CAM definition tool (as the CRT tool is intended for authors with high expertise only). Due to their highly specific character, the virtual reality and simulation authoring tools will not be part of the major evaluation effort within the scope of the project, but later evaluations addressing these tools might be necessary and reasonable.

5.4.2 Method

The target audience of this strand of evaluation will naturally be authors, i.e. interested lecturers, instructional designers and other training providers, who agree to join a special interest group participating in training and evaluation events on the GRAPPLE authoring tools.
As indicated above, two different types of evaluation events with foregoing training are envisaged, workshops and deployments. While the workshops constitute the continuation of the evaluation activities of the formative evaluation phase, the deployments are aiming in realising and stimulating real-world application of the GRAPPL e tools.

In order to allow a direct comparison between subjective usability and acceptance assessments of formative and summative evaluation the same evaluation instruments should be applied as in the already conducted evaluations. For acceptance this would mean the reuse of the questionnaire addressing the different aspects of user acceptance in alignment with the technology acceptance model (Davis, 1986; Jung et al., 2008). With respect to usability in the workshops conducted so far, however, the impression arose that especially the applied standard usability scale IsoMetrics was perceived as rather long. It might therefore be worth considering the use of the much shorter usability measure SUS as suggested for demonstration evaluations, to leverage workload for participants. Though this would mean a change of the evaluation instrument and no direct comparability with the IsoMetrics results, this would allow a comparison to evaluation results from formative evaluation as they were gathered in the context of general level evaluation attempts. Another option would of course be the use of only those subscales of the IsoMetrics that seem especially relevant. This would allow a direct comparison (between formative and summative) of the usability aspects that are of special interest in the context of GRAPPL E. The latter option therefore appears more reasonable. Details on the concrete decisions taken and the questionnaire/subscales to use will be documented in the D8.2b Refined Evaluation Guidelines.

In order to avoid confounding learnability and usability, longer training intervals would be desirable – which would thus make sure that authors are sufficiently trained in the use and functionality of the tools and that subjective ratings on usability are not affected by lacking training (while nevertheless bearing in mind the fact that comparability between the evaluation cycles would be higher and better possible with more similar conditions, e.g. training durations.)

As mentioned above, an extension of training phases is envisaged to be realised by the use of online training measures like Webinars, online courses or training videos etc. Such distance training phases will need to be accompanied by a continuous on-demand support provided to participants in case of arising questions or problems. For an effective training a two-step approach would be conceivable – having an online phase in the beginning, followed by a presence training phase were the tools’ functionalities are presented in more detail and eventual lack of clarity can be overcome. With an efficient supporting system, however, also pure online trainings are conceivable. If the GRAPPL E training material could be implemented and realised in form of an adaptive course, this would provide an additional opportunity of conducting a learner evaluation – with the prospective authors slipping into the role of learners in the course of being trained on GAT.

The presentation of the questionnaires that are going to be reused from the formative evaluation will most likely be done again in form of online survey. This would even allow a responding to the questionnaire on distance in case of pure online trainings. This would however imply that the conditions of filling out the questionnaires cannot be controlled, which is generally the case with online surveys.

In addition to questionnaires, also more qualitative evaluation methods are planned to be used. For an analysis of the conceptual adaptation models created by authors, the thinking-aloud technique can be applied (as in the formative WP9 workshops) in order to get a better understanding of the instructional aims and expectations of authors. In addition to addressing the outcomes of authoring, qualitative methods can also be applied to gather information on the authoring process itself. This shall not primarily serve the elicitation of user feedback on whether authors felt able to appropriately translate their desired instructional strategy into an adaptive storyline. This actually constitutes the authoring perspective on adaptation quality. On principle the use of the same qualitative method in both evaluation strands would be possible. To fit the respective contextual conditions the use of slightly different approaches is suggested.

5.4.2.1 Authoring Workshops

The target audience and also the general procedure for this type of event will be similar to those of the formative evaluation workshops. Participants of evaluation and training workshops will be potential authors that are particularly interested in the GRAPPL E project and system and could imagine making use of such a generic, adaptive learning environment in the future. So, the aim of these workshops is to make users aware of and train them on the functionalities and use of the authoring tools, to allow them to work with the tools, but not the practical application of the tools in educational everyday life.
A main difference between the workshops to those of the formative evaluation phase will most likely be that authors will be enabled to see and check the outcome of a conceptual adaptation model in terms of an adaptive course. Participants will have the opportunity to explore how the adaptive storyline they created is translated by the adaptive engine into a real adaptive course. This is assumed to empower user experience and the authoring process itself and allow corrective actions in authoring.

In general, evaluation workshops shall be conducted at both sites, i.e. university as well as corporate settings. Due to the envisaged procedure of the workshops and especially the suggested increased use of online trainings, this type of evaluation event seems especially suited for business settings.

A focus group during or subsequent to the training phase would be an appropriate way of realising qualitative feedback. Depending on the type of the foregoing training and the concrete evaluation type such a group discussion can be realised as face-to-face or, alternatively, as online focus group. In any case, the focus group will be conducted only after the presentation of the questionnaires, in order to avoid an eventual influence of the group discussion on responses.

5.4.2.2 Authoring Deployments

Deployments are envisaged to address training providers who are ‘early adopters’ and interested in realising and teaching part of their course contents in terms of an adaptive course. This means, that deployments are envisaged to actually constitute first applications of GAT in educational practice. Most probably, the participants will be lecturers from university contexts (i.e. WP9). The involvement of authors from WP10 contexts would be as well desirable but is assumed to be much more challenging.

It is assumed that instructors from different backgrounds and giving different types of lectures will be involved in the deployments. In order to ensure that the instructors gain the necessary skills to define the adaptive courses as they intend with the system, deployments will most likely need a more intensive training and dealing with the authoring tools and thus will by nature involve longer training periods. Also in this case it is conceivable to carry out part of the training in form of online teaching. Nevertheless, deployments will necessarily involve face-to-face meetings for the successful translation of the course contents in question in terms of an adaptive course. The successful realisation of adaptive courses in this case is particularly crucial as the created courses will be the basis for the deployments and according evaluations from a learning perspective (see section 5.5.2.2 Learning Deployments).

Qualitative feedback in this case may also be collected in the scope of focus groups. As the training on the authoring tools will most likely mean a more intensive and possibly even one-to-one contact between participant and trainer/evaluator, also interviews with individual persons are possible. An additional option for gathering information and insight in the authoring process would be the use of the thinking-aloud technique in face-to-face phases. This might be done while working with the authoring tools and trying to build the adaptive course and in this way could be used as an approach of identifying critical events in dealing with the tools (Flanagan, 1954), i.e. incidents that have an important effect on the final outcome.

In addition to the training on GAT, the deployments will also need to include training on GALE for authors as this will be the delivery method of the established courses in the participants’ actual lectures. Therefore, instructors should also have an understanding of the functionality and use of GALE. Although a continuous and on-demand support on the use of GALE is planned to be provided to learners, the instructors should in this way be able to help and respond to issues on short notice.

5.4.2.3 Authoring Tools Targeted

Evaluation of the authoring tools shall on principle be done for the core components of the authoring tools, i.e. the three main tools (DM editor, CRT tool, CAM definition tool). The general training and evaluation events will therefore focus only on the toolset covering the three main tools. Thereby it has to be taken into account that the CRT tool will in general not be used by ordinary authors, but only by authoring experts with high expertise in adaptation and programming skills. Correspondingly, especially DM and CAM definition tool will be the main target of evaluation. For the main tools training and evaluation events will be conducted by different project partners, thus resulting in a large and heterogeneous overall sample of participants distributed over several European countries.

In principle, also evaluations for the specific authoring tools for simulations and for virtual reality would be desirable. As the latter tools, though, constitute rather specific tools, which ordinary authors will normally not need to make use of, they are not in the specified major evaluation scope. One idea for nevertheless realising trainings and evaluations for the simulation and virtual reality tool within the project is to conduct
those events at the partner site’s who have also developed the respective tools and therefore are most skilled in these specific types of authoring processes. The feasibility and significance of such small scale evaluations for the specific authoring tools at the respective developing partners’ institutions is to be discussed and decided upon.

The basic evaluation procedure and methods used will be the same for both, evaluations targeting the three main tools or one of the specific tools. Regarding the questionnaire specifically developed for GAT and therefore covering specific questions or subscales targeting the individual tools, only the relevant questions/subscales will be presented in each case.

5.5 Learning Perspective: Evaluation of GRAPPLLe Learning Environment

5.5.1 Objective

The evaluation of the learning perspective and thus, of adaptive learning experiences with GALE is the designated focus of the summative evaluation phase in GRAPPLLe. At the same time, this perspective of evaluation is also more challenging, in particular as experiences with GALE will be strongly interrelated with the concrete adaptive course used, i.e. the learning content and material and especially also the adaptation strategies. Moreover, the evaluation of the learning perspective shall not only consider short-term learning but also incorporate the aspect of lifelong learning and GRAPPLLe’s potential to support lifelong learning (compare section 4.4 Evaluation of Lifelong Learning).

The general evaluation topics addressed will be usability, acceptance, and adaptation quality, particularly including learning effectiveness. For all those topics, and especially in case of adaptation quality, it needs to be taken into account that the evaluation results may also depend on the learning content and adaptation strategy used. In order to be able to compare and generalise evaluation outcomes different courses with adaptation strategies should be used.

5.5.1.1 Usability

The usability aspect shall investigate the general user experience when working with GALE and give evidence whether GALE is able to realise effective, efficient, and satisfactory adaptive learning experiences. In this regard, especially the comparison between the stand-alone version of GALE to the LMS-integrated versions appears interesting. Differences in usability but also acceptance might be possible due to the fact that user experience of people who are already familiar with a certain learning system can benefit from the LMS-integrated versions. Usability is intended to be assessed by the use of a standard usability questionnaire, like the SUS for a quick and general subjective assessment of usability or the IsoMetrics scale for a more detailed look on different usability aspects.

5.5.1.2 Acceptance

Similar to the authoring tools it is important to address also user acceptance, as even a very good usability of a new technology does not necessarily mean that users accept this technology and intend to use it. Consequently, it is important to elicit also evaluation data concerning aspects that influence users’ acceptance of and intention to use GALE. Similar to the evaluation of GAT, acceptance may be measured by the use of a suitable questionnaire, e.g. querying different acceptance aspects according to the technology acceptance model (Jung et al., 2008). Another option for gathering information on users’ acceptance is in the scope of qualitative methods like focus groups.

5.5.1.3 Adaptation Quality

As elaborated in D8.1a Evaluation Framework (section 4.4) adaptation quality is a comprehensive construct of evaluation involving different aspects that provide information about the benefit of adaptation:

- Learning effectiveness
- Cognitive load
- User feedback on adaptation quality
With respect to learning effectiveness it is important to use also objective measures of learning in addition to subjective ratings of the perceived learning effectiveness. While subjective learning effectiveness can be assessed independent of the particular learning content using the same instrument over different courses, objective measures of learning will require the creation of specific learning tests. In deployment settings on principle also course grades may be used, although they are not able to provide detailed information of learning and will cover learning contents that go beyond those of the adaptive course. Rather, the data of final exams on a course would be useful and would allow sorting out the results for the questions relating to the course part covered with the use of GALE. Such an approach would also allow investigating longer-term learning effectiveness and thus incorporate the lifelong learning aspect. The investigation of the relation between learning effectiveness and other evaluation topics (e.g. cognitive load) would be an interesting option to further investigate adaptive learning. An additional aspect of learning effectiveness under the perspective of lifelong learning is the perceived potential of GALE to support lifelong learning. The investigation of this aspect can be achieved in the scope of qualitative methods.

Cognitive load refers to the effort on cognitive processes put on the learner by the adaptive system. Although this certainly depends to some extent on the course itself, the intention behind this is to investigate eventual impacts of adaptation. On the one hand adaptation shall leverage cognitive load by presenting information relevant and suited for the learner or pointing to the relevant information parts. On the other hand adaptation means changes in the interface, and may therefore in the worst case lead to an increase of cognitive load. Cognitive load can be measured in terms of subjective indices assessed with questionnaires (e.g. NASA TLX; Hart & Staveland, 1988), which has been proven to be a useful approach for instructional research (Paas, Van Merrienboer, & Adam, 1994).

User feedback on adaptation quality refers to the opinion of users whether adaptation was helpful for them and led to an improvement of the system and the learning experience. This can best be queried in the scope of qualitative methods like a group discussion. Of course, this aspect requires that learners are aware of the adaptation which should be ensured in order to allow gathering this type of user feedback. Making the adaptation process scrutable actually also conforms to the users requirements as elicited in the beginning of the project (compare D9.1 and D10.1).

As indicated already in the objectives, the evaluation of the learning perspective will always be strongly related to the particular adaptive course used. This problem is prevalent when addressing the learning effectiveness of a generic learning system that is on principle intended to be open to any knowledge domain, course, and pedagogical strategy. Naturally, adaptation quality does not focus on the quality of the underlying technology, as it is the case when considering usability and acceptance aspects. Rather, the benefit of adaptive learning is investigated, and thus, to a certain degree, the quality of the course design and designer. The adaptation process is defined by the authoring expert, only if this definition is done in a sound way, learners may benefit from the adaptive course. Correspondingly, evaluation results for a certain course provide only an indication on the effectiveness of adaptivity. If positive effects for a variety of adaptive courses can be shown, the benefits of adaptive learning can be demonstrated more convincingly – always under the condition of a successful and sound authoring. This is ensured to a certain extent by documentation and training on GAT, on the one hand, and by the assumption of the availability of appropriate didactical skills in authors and instructors, on the other hand.

5.5.2 Method

Also in case of the learning perspective evaluation will be coupled with training events. The aim of the training events will be to make learners familiar with what adaptivity means, the functionalities provided by GALE, and the possibility to experience and deal with adaptive course material. Learners possibly are even less familiar with adaptive systems in general. The provision of an understanding for adaptation is important in order to convey the knowledge necessary for giving a judgement about adaptation quality.

Similar to the GAT evaluation also the evaluation from a learning perspective is suggested to be realised in a two-fold way – carrying out workshops and deployments. Workshops will provide the opportunity to realise systematic comparison studies with and on GALE, and deployments will allow investigating the significance and applicability of GALE in educational practice.

5.5.2.1 Learning Workshops

Target audience for these workshops will be small groups of users that can be regarded potential future learners dealing with the GRAPPLE learning system. These may be learners who are interested in adaptive learning environments and/or students or employees of training providers envisaging future use of GRAPPLE facilities. The learning workshops can be seen independent of the authoring workshops, i.e. the
adaptive course examples created in the authoring workshops will not be the basis for the learning workshops.

Such learning workshop will consist in a training on adaptivity, the functionality and use of GALE and a subsequent learning session, where participants are asked to work through an adaptive learning course. Similar to authoring workshops this training may be conducted face-to-face or online. While face-to-face phases allow a more intensive training and responsiveness to learners’ needs and questions, online phases would particularly bring the benefit of greater flexibility. To get the best of both, a combination of face-to-face and online phases is recommended.

The training and learning part of the workshop would then be followed by the evaluation phase, collecting data on usability, acceptance, and adaptation quality. In contrast to the learner evaluations carried out in the formative evaluation of WP10, in this case a real learning session aiming at acquiring knowledge will be realised. Consequently, learning effectiveness can be assessed not only in terms of subjective but also in terms of objective measures. Objective learning effectiveness will in this case rather concentrate on short term effects.

For the conduction of learning workshops the course contents can be kept constant over different events and partners/institutions and therefore would allow systematic comparison in order to investigate the particular benefit of adaptive learning. A knowledge domain that is equally appropriate for different audiences and does not require any particular pre-knowledge or expertise should be chosen for this purpose, for instance the solar system or Milky Way. The adaptive course(s) used for the learning workshops would be created by the GRAPPLE consortium, probably reusing existing adaptive course material.

The learning workshop appears especially suitable for conducting comparison studies, i.e. systematically varying certain factors while keeping constant the other conditions. One possibility would be the comparison between different adaptation strategies; courses with different strategies or techniques (e.g. adaptive annotation vs. link hiding) could be used within each workshop or across different events.

In order to demonstrate the effectiveness and added value of adaptive learning, it is also important to realise a comparison between adaptive (as experimental group) and non-adaptive (as control group) learning experiences. In this case, however, those learners receiving the non-adaptive course might be frustrated or disappointed when first learning about what adaptivity and afterwards being in the non-adaptive condition. Therefore, the training of the control group should be slightly different and confined to non-adaptive functionalities of the learning system. For this reason, a comparative study of this type appears rather complicated in the scope of deployments. Moreover, the systematic comparison of adaptivity vs. no adaptivity should be realised under conditions as much controlled as possible. Comparisons between adaptive and non-adaptive versions should be realised in terms of presenting courses by the use of an LMS integrated with GALE vs. the LMS alone.

As elaborated in the first version of the evaluation framework (D8.1a, sections 3.3 Evaluation of Adaptive Systems and 5.5.1 Adaptation Layers) the evaluation of the GRAPPLE learning system should optimally also realise approaches of layered evaluation. This would allow a sound investigation of quality of adaptation by considering individual components of the adaptation process. To this end, the layered evaluation model of Brusilovsky, Karagiannidis, and Sampson (2004) shall be adopted, which distinguishes two main and distinct processes of adaptation:

1. **User modelling**
   - The goal of this phase of adaptation is to draw inferences from different aspects of user-computer interaction to user characteristics. User modelling is usually carried out on the basis of low-level information provided by a system’s monitoring mechanisms (e.g. task completion, answers to test or quizzes, keystrokes). The results of the user modelling process are stored and represented in the user model, which captures information on user characteristics that are considered important for a particular application and that are used for the purpose of adaptation.
   - When evaluating user modelling, the question is whether user characteristics are successfully inferred by the system and stored in the user model. Comparisons of the opinion of experts monitoring users or of users themselves with the conclusions made by the system may serve this purpose.

2. **Adaptation decision making**
   - The adaptation process itself constitutes the adaptation decision making, where a specific adaptation is selected and grounds on the results of the first phase, i.e. user modelling. The rationale and logic behind this decision process is captured by adaptation rules determining which constituents of adaptation are selected according to the results of user modelling.
• When evaluating this layer the question is whether adaptation decisions are appropriate and meaningful given a certain user model constellation. This question can be addressed by direct knowledge testing (in case of knowledge-based adaptation) or by querying the user on adaptation quality (e.g. whether adaptation improved the quality of the system given a certain goal).

A possible design for an evaluation study separately investigating these two adaptation layers has been elaborated already in D8.1a (section 7.5 Learner Evaluations). Such a comparative study may be carried out in the scope of learning workshops. This kind of study however requires more effort in controlling and realising the different experimental conditions. As a consequence, most probably only part of the workshops will be able to implement this design and only part of the partners involved in evaluation will be able to realise the design. This design assumes an adaptive course using knowledge-based adaptation, e.g. using adaptive annotation. For sufficiently evaluating the adaptation layers three conditions should be realised, two experimental and one control condition:

- Ex1: Tailored adaptation
- Ex2: Arbitrary adaptation
- Contr: No adaptation

This means, when trying to address separate adaptation layers one additional experimental group (Ex2) needs to be realised compared to a pure comparison of adaptive with non-adaptive learning. For the two experimental groups matched pairs should be realised, i.e. creating pairs of participants which have similar characteristics (e.g. gender). The experimental condition ‘tailored adaptation’ would realise full adaptive functionality, i.e. learners receive adaptation according to their knowledge as inferred by the system. The second group, ‘arbitrary adaptation’ will receive adaptation, which however is not tailored to the learner’s knowledge but is rather more or less arbitrary. This is realised in terms of a yoked-control design (e.g. Kickmeier-Rust, Marte, Linek, Lalonde, & Albert, 2008) by providing each member of this group the exactly same adaptation behaviour as the other member of the matched pair from the Ex1 group has received. Consequently, the participants of the Ex2 group receive adaptation, but not tailored to their personal knowledge. The third group will be presented with the learning content without adaptation, while still realising user modelling in the background. A visualisation of the design and the evaluation aspects and hypotheses addressed in this kind of evaluation study is represented in Figure 17.

For investigating the adaptation layer of user modelling, the results on a learning test assessing objective learning effectiveness are compared with results of user modelling. If user modelling is successful then the knowledge inferred by the system for a learner should correspond to the objective performance in the learning test. This can be investigated for the overall group of learners, i.e. over all groups. If user modelling cannot be realised in the non-adaptive version (e.g. because of using the LMS alone without any GALE functionality integrated), this may also be confined to the experimental groups (compare Figure 17).

For investigating adaptation decision making, the results on adaptation quality of the different conditions are contrasted. Learning effectiveness is assumed to be significantly higher for learners receiving tailored adaptation than for the other two groups of learners. Furthermore, cognitive load is assumed to be
significantly lower for tailored adaptation than for the other two groups. When contrasting user feedback on adaptation quality for the two experimental groups results should be better for the tailored adaptation group than for the arbitrary adaptation (compare Figure 17).

The suggested design allows an appropriate comparison between different groups. The two experimental groups would feature equivalent environmental conditions (i.e. both members of a matched pair receive the same adaptation). This allows evaluating whether the presentation of tailored adaptation is more effective than when receiving no adaptation by isolating effects of adaptation separately from other possible factors that may influence the effectiveness of adaptation, such as the number of changes due to adaptation. Moreover, the comparison of an adaptive version of a system with a non-adaptive one (Contr) that is realised by the use of the LMS-integrated version of GRAPPLE vs. the LMS alone is considered to be a fair comparison.

In addition to the consideration of the different aspects of adaptation quality, it would be interesting to compare also results for usability and acceptance collected in the different conditions. This does not serve the evaluation of the individual adaptation layers but rather only the comparison of the different versions of the system (i.e. conditions). Optimal, the results on these evaluation topics should also be the best for the version realising tailored adaptation.

This kind of experimental design would, however, require the implementation of a record-and-play functionality which allows the realisation of the experimental condition Ex2 (arbitrary adaptation). It has therefore to be clarified whether the proposed design can technically be realised. In particular, the significance of this evaluation approach also needs to be discussed in light of the mentioned problematic of the dependence of evaluation outcomes on the specific adaptive course used. Consequently, the quality of user modelling as well as adaptation decision making will always be linked to the quality of the underlying user model and adaptation rules defined by the author. A targeted evaluation of the different adaptation layers therefore seems especially important for an adaptive learning system or even an educational game with a fixed course content and adaptation strategy. For a generic system like GRAPPLE, on the contrary, the value of such an evaluation approach is more questionable. It therefore might be more valuable for GRAPPELEvaluation purposes to try to demonstrate the overall potential of adaptive learning by evaluating adaptive learning and its learning effectiveness and other adaptation quality aspects on a more general level.

Evaluation instruments in the workshops will consist of questionnaires, especially for usability and cognitive load. For objective learning effectiveness a learning test on the course contents will have to be created. Similar to the GAT evaluation, further qualitative methods are foreseen in order to get a better understanding of the processes of learning with GALE. In order to understand better the learning processes and learners’ experiences and attitudes towards the system, including aspects of acceptance and adaptation quality, focus groups (face-to-face or online) would be suitable. Moreover, the learners’ opinion on whether GALE is (suitable) to foster lifelong learning could be queried in this way. For the deeper consideration of the interaction processes while learning with GALE the think-aloud technique appears appropriate and would also allow to identify critical events or incidents in the use of GALE. The think-aloud method would require face-to-face situations for the learning phases with individual learners and will generally only be applicable in workshop settings but not in deployments.

5.5.2.2 Learning Deployments

The learning deployments will build upon the authoring deployments and are directly related to them. Actually, they constitute a continuation of the deployment in the respective settings; the adaptive courses created in the authoring deployments will be used in the learning deployments. The focus of the learning deployments is the applicability and added value of GALE in educational practice. In particular learning effectiveness shall be considered, also in terms of longer-term learning effects.

As sketched already in section 5.4.2.2 Authoring Deployments the idea is to realise part of an actual course or lecture in terms of an adaptive course. Most probably, this will be possible in university settings (WP9), although of course the involvement of corporate settings (WP10) would be desirable. Target audience of these deployments are the students of those instructors who have taken part in the authoring deployments.

Learning deployments will also involve training measures in order to convey knowledge about adaptivity and GALE functionalities and use. The conduction of this training will most probably be done by the evaluator, but is also conceivable to be taken over by the lecturer. Subsequent to the training, the adaptive courses created by the instructors can be provided to the learners as part of the overall course/lecture. During the learning phase a continuously and immediately available support system should be provided to learners. This shall serve overcoming probable technical issues or problems in the use of GALE and thus, avoiding confounding learnability with usability or other aspects of user feedback.
A main difference to the learning workshops will be that the learning contents and general conditions of training and evaluation will not be constant over different deployments. As adaptive courses created by different instructors from different backgrounds and institutions and giving different types of lectures will be used for the learning deployments, there will be a large variety in the knowledge domains covered as well as in the adaptation strategies realised in the courses. If certain benefits (or drawbacks) of adaptive learning can be identified across this diversity, these outcomes would be generalisable to adaptive learning experiences with GALE.

The adaptive learning experiences in the learning deployments will generally be longer and more intensive than those carried out in learning workshops. This will allow the examination of the different evaluation topics after a longer-term use of and interaction with GALE in educational practice. This refers to usability and acceptance, but also to learning effectiveness. When using immediate learning tests and longer-term knowledge assessment, for instance through final exams, short-term and long-term effectiveness could be investigated. To some extent, adaptive learning and traditional learning could also be contrasted, by comparing learning effects on contents covered by the adaptive course with those covered by the remaining lecture. Hereby, differences in different subtopics covered as well as eventual recency effects of learning contents would need to be taken into account. A more appropriate comparison would be to compare parallel courses, one including the use of GALE and one not using it. As however, the deployments focus on applicability of GALE and systematic comparisons are rather in the scope of the workshops, this aspect will likely not be part of the concrete evaluation methods that will be specified in detail in the evaluation guidelines (D8.2b).

Learning effectiveness should optimally be assessed not (only) directly after the learning sessions. On the contrary, evaluation data on usability, user acceptance, cognitive load, and user feedback on adaptation quality should be collected promptly after the actual experience with GALE in order to avoid any subsequent shifts or effects on users’ opinions or memory.

Evaluation instruments in the deployments will be similar to those used in the workshops. Questionnaires will especially target usability and cognitive load. For objective learning effectiveness learning tests on the course contents covered may be generated for the purpose of evaluation. Final exams may also serve for learning outcome measurement, as long as the covered course contents are appropriately reflected and identifiable in the exam. For the collection of qualitative feedback regarding acceptance, adaptation quality, and general user experience again the conduction of focus groups – with those learners who agree to participate – would be appropriate. Group discussions in the learning deployments will especially also be able to cover the aspect and question whether GALE is perceived to be able to support lifelong learning. In order to realise a smooth and uncomplicated administration of the evaluation instruments and methods and flexible feedback opportunities, the use of online questionnaires and online focus groups seems valuable.

5.5.2.3 Versions of GALE

For the summative evaluation cycle it is planned to include in addition to the stand-alone version of GRAPPLE also the system versions integrated with the different LMSs. The versions with LMS are actually important for the evaluation of usability, acceptance, and learning effectiveness with learners. This will enable on the one hand the comparison between the adaptive version of a system with a non-adaptive one (i.e. LMS with and without GRAPPLE functionalities), which is an appropriate comparison for investigating learning effectiveness. Furthermore, this will also allow the comparison of the stand-alone version with the LMS-integrated versions (i.e. adaptation with LMS vs. without LMS), which is interesting from the perspective of usability and user acceptance.

For comparing stand-alone and LMS-integrated versions it is suggested that partners who are involved in the integration of GRAPPLE with a certain open-source or commercial LMS, carry out their training and evaluation events by the use of the respective integrated version. In addition, if in certain deployments a particular LMS is commonly used (e.g. use of Moodle as standard LMS at a university), this would also argue for the use of the respective integrated system. Evaluations carried out by partners with no in-depth experience with a certain LMS and with no particular contextually given LMS experiences, are suitable for the use of the stand-alone versions. In both scenarios also the opportunity of comparing LMS-integrated and stand-alone version would be an option.

For the LMS-integrated versions of GALE in any case prior experiences with the respective LMS need to be queried in order to analyse the user ratings in dependence of existing or no experience with the LMS. As it can be assumed that prior experiences with an LMS might influence the user experience with GALE this needs to be taken into account when analysing the collected evaluation data.
It has to be noted here that the combination of GALE with an LMS also raises some difficulties for evaluation. This is because the main aim of a proper integration of LMS and the GRAPPLE system is that the user is not aware that some of the used services are offered by the LMS and others by the GRAPPLE components – they shouldn’t even be aware that they are facing a learning environment that is made up of the marriage of two different systems (DeBra, Smits, Pechenizkiy, & Vasilyeva, 2008). As a consequence, the evaluation results (e.g. on usability) of the learning environment as a whole will actually be mixing up the effects of both. The conclusions drawn from such an evaluation cannot be clearly and singularly be associated with the GRAPPLE system as such, but may be biased by or highly related to the used LMS. Therefore, the evaluation of the stand-alone version of the system is important in order to be able to clearly attribute assessed effects to the technology developed in the GRAPPLE project.

5.6 Evaluation of Visualisation Tools

5.6.1 Objective
The main objective of summative evaluation attempts on the visualisation tools is the investigation of the visualisation benefit of the actual implementations of these tools. While the formative evaluation was carried out only with visualisation mock-ups and had the primarily developmental purpose of identifying weaknesses and issues for further improvement, the summative evaluation with the implemented visualisations is intended to demonstrate the benefit of the visualisations and to give a concluding statement of their quality.

The overarching evaluation question is whether the visualisation tools serve their intended purpose. This relates to the visualisation quality and expected benefits for users. Visualisations should in general enhance meta-cognition and leverage cognitive load, which can be regarded as two general aspects of visualisation quality. For learner visualisations an empowerment of learning effectiveness and collaboration between peers would be expected. For instructor visualisations an improvement of teaching by making aware of different needs of learners and facilitating individualisation would be expected. In addition, usability (aspects ‘suitability for the task’ and ‘self-descriptiveness’) and acceptance of the visualisation tools should be investigated.

5.6.2 Method
Target audience of these evaluations will be potential future end users of the GRAPPLE system. This should involve training providers as well as learners, as the visualisation tools are intended for both stakeholder groups. Thereby, the visualisations intended for instructors should be evaluated by training providers and visualisations for learners should be evaluated by learners. In each case, appropriate measures of training need to be realised before the evaluation, i.e. introducing the visualisations, explaining their purpose, the information depicted.

The evaluations of the visualization tools might be done in special evaluation studies or in the scope of the in-depth evaluations for GALE. The evaluation of instructor visualisations requires also training measures on GALE, where the visualisations are embedded. This is rather the case in the authoring deployments, where instructors are also introduced to GALE and will likely be interested in monitoring the learning progress of his/her students by the use of the instructor visualisations. Learning workshops and deployments, on the other hand, would enable the evaluation of learner visualisations. For the realisation of systematic comparison studies on the visualisations special evaluation workshops should be conducted, in order to avoid having too complex evaluation designs with too many conditions (as it probably would be the case if including it into comparison studies of learning workshops). One option would be that the partner(s) who has (have) been mainly involved in the task of implementing the respective visualizations conduct(s) learning workshops that systematically vary conditions on visualizations while keeping the other conditions constant.

To investigate the benefits of visualisations for learners not only subjective measures as used in formative evaluation should be used, they need to be complemented by objective measurements e.g. of learning effectiveness. When evaluating the benefits of the visualisations between-subjects designs should be realised with the conditions ‘visualisation present’ vs. ‘visualisation absent’. One group of user would then work with the GRAPPLE system while having available the visualisation tools (experimental group) and one group of users would not be provided with the visualisations (control group). For the control group no training on the visualisations should be realised in order to avoid frustration of not being provided with the visualisations.
In such a comparative study, benefits for learning and collaboration in case of learner visualisations could be investigated, comparing objective learning measures and user feedback on the extent and easiness of collaboration for the two conditions. The hypothesis in this case would be that the use of visualisations leads to better results with respect to learning as well as collaboration. For both types of visualisation meta-cognition could be compared, assuming that the use of visualisations leads to improvements in meta-cognitive abilities and processes. Meta-cognition in terms of self-realism or knowledge awareness can be investigated by contrasting learners’ subjective learning performance with their actual, objective learning outcomes. The experimental group using the visualisation should therefore show a better conformance between objective and subjective learning outcomes than a control group without the visualisation. Information on meta-cognition referring to instructor visualisations can be collected in terms of subjective assessments. Additionally, visualisations are assumed to reduce cognitive load. This can be investigated by comparing subjective measures of cognitive load of the experimental group with the control group.

In addition to these systematic comparisons evaluations should also gather user acceptance and usability issues, as well as qualitative user feedback (w.r.t. visualisation quality). The evaluation instruments thereby need to explicitly refer to the visualisations, in order not to confound the evaluation outcome with the overall GALE. (Conversely, as the visualisations are an integral part of GALE the evaluations of GALE need not explicitly factor out the visualisations from users’ subjective assessments.) Relevant usability aspects may be assessed by a short questionnaire, e.g., adopting the respective items from the questionnaire used in formative evaluation (see section 2.2.3). For feedback on acceptance, perceived visualisation quality, and processes/problems of dealing with the visualisations qualitative methods (focus group) can be applied.

6 Conclusion and Next Steps

This document has presented the second and final version of the GRAPPLE evaluation framework. It builds upon and complements the first versions of the evaluation methodology documents, which have elaborated in detail on the general evaluation objectives, topics, and scope in GRAPPLE. The present document augments the methodology by further considerations and descriptions on the summative evaluation aspect.

In the beginning of this document, addenda and adjustments to the formative evaluation methodology are presented. These have been done in close cooperation with WP9 and WP10 in order to correspond to the requisites of the empirical settings of university and business. Reflections on the methodology towards the final evaluation cycle are made. Subsequently, the evaluation methodology for the summative evaluation phase is focused. To this end, relevant methodological issues and state of the art are discussed. On this basis, the methodology for summative evaluation in GRAPPLE is elaborated on a conceptual basis, arguing again for an alignment of training and evaluation events. Summative evaluation attempts shall continue the investigation of the usability and acceptance of GAT, but focus especially on the investigation of adaptive learning experiences with GALE. An evaluation scheme is presented that foresees different levels and types of evaluation, in order to ensure a comprehensive investigation of the quality of the GRAPPLE tools and facilities from different perspectives. The objectives and possible methods of the different levels and types are presented, providing a general idea and pool of considerations as a basis for the specification of the concrete evaluation procedures. The summative training and evaluation sessions are planned to be realized to a larger extent also by the use of virtual environments, bringing in this way greater flexibility and outreach of GRAPPLE’s training and evaluation.

In a next step, the elaboration of the summative evaluation methodology will be continued towards translating the methodological framework presented in this document into more concrete designs and procedures for the actual conduction of methodologically sound empirical evaluations in the WP9 and WP10 settings. Evaluations fully following the considerations presented in this document will likely go beyond the scope of the project itself. For summative evaluation it is necessary to identify what realistically can be done within the project lifetime, also from the viewpoint of technical feasibility, and which aspects may have to be shifted to later evaluation phases. The concrete planning for the final evaluation cycle is done in the scope of the Evaluation Guidelines task and deliverable (D8.2b). In particular, the considerations on evaluation workshops and deployments as outlined in the present document need to be elaborated further. These considerations are rather ambitious and need to be substantiated in terms of feasible and tangible evaluation plans for the final phase of the project. This will mean to clearly define the focus and scope of the actual training and evaluations to be accomplished in the individual settings and the responsibility of the partners involved in evaluation.

The guidelines will detail the evaluation instruments (e.g., questionnaires, moderation guides for focus groups etc.) and procedures for data collection and analysis in the evaluation studies. In this way, they serve as the evaluation manual for the empirical work in WP9 and WP10 and provide project partners involved in the
empirical evaluations in higher education applications and industrial outreach with support and guidance in their empirical work with users.

The procedures and stages of evaluation need to be harmonised with the training plans and activities of WP9 and 10, as an appropriate foregoing training on the GRAPPLE tools constitutes a necessary prerequisite for evaluation. In particular, the evaluation guidelines also need to elaborate the concrete methodology by responding to the contextual conditions of the respective empirical settings. As a result, the evaluation guidelines will be specified on the basis of cooperation and exchange with WP9 and WP10.

A critical prerequisite for accomplishing empirical training and evaluations is of course the availability of the software components and tools that are to be evaluated. The final implementations of the GRAPPLE tools and their integration including testing need to be accomplished in a timely manner. Only if the tools are available and ready in an appropriate and mature status of implementation, training and evaluation events can be organised and carried out in due time. The integration of the GRAPPLE components is particularly critical for the deployments. Deployments are envisaged as first applications of the GRAPPLE environment in educational practice, which is only possible with a smoothly working, complete system. Otherwise, it would probably be better to realise deployments following a different strategy first. This could consist e.g. in approaching users in the educational field and practice and demonstrating them with the GRAPPLE environment and objectives. In this way, interest in the GRAPPLE tools could be provoked and the notice, use, and spread of adaptive learning could stimulated.

The summative evaluation methodology developed in the tasks and documented in the deliverables of WP8 on a theoretical basis, will be put into evaluation practice in university and corporate settings in order to gather empirical data and results on the benefit of the GRAPPLE tools, users perception and attitude towards them, and in order to demonstrate the overall achievement of the project.

References


Appendix: Further Information on the Qualitative Evaluation Methods

The following subsections present further information and state of the art on the qualitative evaluation methods presented in section 4.3.

A1 Think-Aloud Method

A1.1 The Think-Aloud Method in E-Learning Evaluation

The think aloud method has been controversially discussed in the area of usability research for more than twenty years (Ericsson & Simon, 1984; Nielsen, Clemmensen, & Yssing, 2002). Although there are some drawbacks, like the guidance given to participants, observer influence, and the complexity of data analysis, the richness of the data collected is able to outweigh these constraints and therefore the think-aloud method has a high potential and value in the field of e-learning evaluation (Cotton & Gresty, 2005). It has been proven to provide reliable results on explicit intentional behaviour and expectations of people towards technical system designs (Wright & Monk, 1991; Boren & Ramey, 2000).

Think aloud protocols are verbalisations of intentional activities including the related intentions, expectations and – if applicable – related concepts. These verbalisations are captured while a person performs the activity of interest. Instead of simply performing the activity the person reports about the rationale behind the action while performing the action. In case of unexpected developments (e.g. when a computer system throws an error message) the person also reports what was expected. This procedure allows capturing sequences of verbalisations that allow the reproduction the process and the related subjective intentions. Different to the origin of think-aloud protocols as an experimental instrument of cognitive psychology that analyses similar behaviour; the application of think aloud protocols in usability research focuses on highly variable behaviour including highly unexpected behaviour in order to identify strengths and weaknesses of a tool or a procedure. Furthermore, think aloud protocols are also used in usability research for capturing verbalisations of participant preferences simultaneously with performance data (Boren & Ramey, 2000).

The main benefit of this qualitative approach as part of a formative evaluation is that the findings can be almost directly used as use-case definitions of the unified modelling process (UML; Arlow & Neustadt, 2005), because each think-aloud protocol reflects the intentional behaviour and related expectations of an application user.

A1.2 Strengths and Weaknesses of the Think-Aloud Method

In the scientific communities of psychology and computer science the think-aloud method plays an important role. It is often described as one of the most widely used evaluation methods, especially for testing the usability of different computer interfaces.

Van Velsen, van der Geest and Klaasen (2007) elaborated on the strengths and weaknesses of the think-aloud method in a study testing the usability of a personalised system by comparing the use of thinking-aloud with the use of interviews and questionnaires. Qualitative evaluation approaches like the think-aloud method are by some researchers considered more useful than quantitative evaluation approaches like questionnaires (Gena, 2005). The think aloud method allows to understand the way user really interact with a system, and appreciate personalised systems. Therefore the Van Velsen et al. (2007) explored in their study what kind of (unique) issues interviews, questionnaires and thinking-aloud are able to elicit when used for evaluating a personalised system. These issues refer according to Jameson (2006) to: predictability, comprehensibility, unobtrusiveness, privacy, breadth of experience, system competence. Results showed that thinking-aloud elicits more negative comments from participants than interviews and questionnaires. Questionnaires elicit more comments on some usability issues and interviews elicit more comments on most usability issues and appreciation of personalisation compared to thinking-aloud. Thinking-aloud focuses on more detailed issues and so it’s possible to elicit more comments from participant on perceived relevance of the information presented by the system than interviews and questionnaires. Thinking-aloud uncovers a unique set of issues with the system. However a large part of issues demonstrated by thinking-aloud were also detected by questionnaires or interviews. The outcomes of this study revealed that there is one topic of feedback for which thinking-aloud was the best elicitation technique: perceived relevance of search results. This may be the most important condition for user acceptance; so think-aloud should be an important part of the iterative design process.

Smilowitz Darnell, and Benson (1994) also illustrates the strengths and weaknesses of the think-aloud method. In his work he used think-aloud in a lab setting, take-home questionnaires for beta testing and a
discussion forum on an intranet-based bulletin board system to gather user feedback. Results showed that the lab tests were significantly better in uncovering usability problems than the forum tests. The beta and lab testing were equally effective, but by using the thinking-aloud method in the lab many more high severity problems could be uncovered.

The limitations or problems of think-aloud data refer to the reactivity of the participants, participant’s verbal abilities, and the validity of data (Young, 2005). Moreover, as it is the case with qualitative methods in general, data analysis of think-aloud data is by nature complex and effortful than it is the case with quantitative data collection techniques.

There are three main effects of asking a participant to think aloud, which have to be taken into account when applying this method. The first relates to the ability of a person to think aloud while to be engaged in a task at the same time. The second considers the effects of talking aloud during an activity which would normally perform in silence. The last one is the effect of drawing a participant’s attention to the cognitive processes underlying the task being undertaken. Participants generally vary in their ability to produce useful think-aloud data. Furthermore, only a small subset of information – the contents of working memory – is accessible and available to consciousness and so the data produced is automatically limited (Wade, Buxton, & Kelly, 1999).

To sum up, actually performing thinking aloud might be somewhat uncommon or awkward for users – it may influence the way people are actually performing and some may find it hard to talk when concentrating on the problem/task. Still, thinking-aloud is a valuable and widely used evaluation method; when asking users to say what they are thinking and doing insights into what the user is thinking and expecting can be gained.

A1.3 Data Analysis of Think-Alouds

Data collected by verbalisation can be classified into “soft” data and “hard” data. “Soft” data are obtained when data interpretation occurs simultaneously with the verbalisations. The interpretation of these verbalisation data is liable to the theoretical presuppositions of the interpreter (Ericsson & Simons, 1984). “Hard” data refers to the case when the verbalization of the person is recorded and then transcribed. Ericsson and Simons (1993) propose to make verbatim transcripts of the recorded tapes after collecting the verbal data; this allows obtaining the raw data in a “hard” form, which can then be analyzed objectively.

To interpret the “hard” data given by the subjects, the verbalisation needs to be encoded. To achieve an objective analysis of the subjects’ behaviour theoretical presuppositions have to be considered. First of all the inferences and expectations of the person developing the coding scheme have to be minimised. Furthermore, persons doing the actual coding of the protocols should not be involved in the development of the coding scheme. Finally, the protocols should be transcribed by at least two people who work independently. The results of the protocols can then be compared and any differences can be resolved.

Payne (1994) defines two different approaches to coding data:

A. Code instances in which certain types of thought seem to occur within a protocol. The frequency of occurrence of different types of reasoning can then be computed across problem types or individuals;

B. The protocols are broken up into short phrases or segments. Each phrase should refer to what constitutes as single task assertion or reference by the subject. These segments can then be coded and analysed.

To facilitate the process of transcription several computer-based systems are available like ‘Gesprächsanalytisches Transkriptionssystem’ (GAT), ‘halb-interpretative Arbeits-Transkription’ (HIAT), or Codes for the Human Analysis of Transcripts (CHAT).

B1 Focus Groups

B1.1 Focus Groups compared to Surveys and Individual Interviews

In the literature studies can be found which include the comparison between focus groups and other, traditional methods such as surveys.

In Ward, Bertrand, and Brown (1991) compared focus group and survey results from three studies on family planning. The found results of this comparison were a high similarity for 30% of the variables, focus groups provided more information for 42% of the variables, surveys provided more information for 17%, and there is dissimilarity for 12% of the variables. The biggest difference found between the methods was the ability of the focus groups to produce more in-depth information on the topic at hand.
Another comparison by Saint-Germain, Bassford, and Montana (1993) showed that the findings from the focus group interviews in most cases could confirm the findings of the previous population surveys. In many cases, though, the focus group interviews went even beyond the information obtained in the survey, thus amplifying the understanding of the.

Morgan (1996) summarises several comparison studies. He describes a consistent set of difference: The survey interview setting limits what respondents say about sensitive topics, in comparison to what they are willing to reveal in focus groups. The differences in response options mean that surveys are in general better able to elicit yes/no answers about specific behaviours and experiences, even though the forced-choice format of the survey items limited what respondents could say on general attitude areas, in comparison to the more open-ended discussions in the focus groups.

Ward et al (1991) explicitly noted that all of these comparisons used only the variables that occurred in both studies, thus downplaying the fact that the surveys typically covered many more topics than did the focus groups. There is thus a key tradeoff between the depth that focus groups provide and the breadth that surveys offer.

Wight (1994) compared individual interviews and focus groups by focusing not on the generated number of ideas but on the comparability of the results they produce. The study involved both group and individual interviews with the same participants. Adolescent males were interviewed concerning their sexual experiences. The sequence of the two types of interviews was systematically varied. Wight reported that the research participants are more likely to express macho attitudes (with a male researcher) or to sexually harass (a female researcher) in group setting than in individual interviews in which they often use gender-sensitive terms. The results showed a high discrepancy between reports of participants who participated in individual interviews first and then in focus groups. Participants who started in focus groups gave similar accounts in the individual interviews.

**B1.2 Strengths and Weaknesses of the Focus Group Method**

Steward (2007) summarises the strengths and the weaknesses of the focus group method.

**Strengths**

- Compared with a separate interview of each individual gathering data from a group of people is much more quickly done and often at less cost.

- Focus groups allow the researcher to interact directly with the respondents. The clarification of responses, follow-up questions, and the probing of responses are provided. Furthermore the researcher has the opportunity to observe nonverbal responses such as gestures, or smiles.

- Due to open response format of a focus group large and rich amounts of data in the participants own words can be gathered.

- In a focus group participants react to and build on the responses of other group members. The so emerged group dynamic provides data or ideas that might not have been uncovered in individual interviews.

- Focus groups can be used to examine a wide range of topics with a variety of individuals and in a variety of settings.

**Weaknesses**

- The interaction of respondents with one another may have two undesirable effects. At the one hand the responses of group members are not independent of one another. At the other hand the results obtained in a focus group may be biased by a very dominant or opinionated member. More reserved group members may be hesitant to talk.

- Summarization and interpretation of results may be difficult because of the open-ended nature of responses.

- The moderator may bias results by knowingly or unknowingly providing cues about what types of responses and answers are desirable or seeking to achieve group consensus on particular topics.

**B1.3 Procedure and the Level of Moderator Involvement**

The ideal group size for realising a focus group is between four and eight people. Lasting one to two hours the session should be comfortable and relaxed. Ideally the group discussion should be audio taped and transcribed.
The role of a moderator or group facilitator is one of the most significant features of focus groups. Depending on the moderator's degree of control, Morgan (1992) distinguishes between “more structured” discussion and “less structured” discussion. With regard to a more structured group discussion, there are two aspects which have an influence:

- First, it can be more structured with respect to asking questions. The facilitator controls what topics are discussed (e.g., moving things forward when the conversation is drifting or directing attention away from less important issues by steering the conversation back on course.

- Second, it can be more structured with respect to managing group dynamics. The interacting of the participants is controlled by the moderator by e.g. trying to get everyone to participate equally in the discussion.

Regarding less structured discussion, the facilitator's involvement asking questions or managing group dynamics is slight. The group can pursue its own interests and is allowed to talk as much or as little as they want.

According to McDonald (1993) who evaluated differences in moderator style, approaches to moderating should be linked to research goals. It implies that arguments about whether moderators should use a more or less structured approach are meaningless unless one specifies the goals of the research.

**B1.4 Data analysis**

Analyzing the data gathered in the scope of focus groups is basically the same as analyzing any other qualitative self-report data. Data is recorded either by taking notes, or, preferably (or additionally) by recording the discussion, which allows a more complete and objective collection of data. The gathered data is then transcribed verbatim (word for word). After this, the data has to be organized by creating different coding categories, such that it is possible to code and sort the gathered data (compare also A1.3).