Note: This is a preprint. The final published version of this article is available from Springer at http://dx.doi.org/10.1007/s11257-010-9082-4

Layered Evaluation of Interactive Adaptive Systems: Framework and Formative Methods

Alexandros Paramythis
Institute for Information Processing and Microprocessor Technology (FIM)
Johannes Kepler University Linz
Altenbergerstr. 69
A-4040 Linz, AUSTRIA
+43 732 2468 8442
alpar at fim.uni-linz.ac.at
http://www.fim.uni-linz.ac.at/staff/paramythis/

Stephan Weibelzahl
National College of Ireland Dublin
Mayor Street IFSC
Dublin 1
Ireland
+353 1 4498 579
sweibelzahl at ncirl.ie
http://www.weibelzahl.de/

Judith Masthoff
University of Aberdeen
Aberdeen AB24 3UE
Scotland, UK
+44 1224 272299
j.masthoff at abdn.ac.uk
http://www.csd.abdn.ac.uk/~jmasthof

Abstract: The evaluation of interactive adaptive systems has long been acknowledged to be a complicated and demanding endeavour. Some promising approaches in the recent past have attempted tackling the problem of evaluating adaptivity by "decomposing" and evaluating it in a "piece-wise" manner. Separating the evaluation of different aspects can help to identify problems in the adaptation process. This paper presents a framework that can be used to guide the "layered" evaluation of adaptive systems, and a set of formative methods that have been tailored or specially developed for the evaluation of adaptivity. The proposed framework unifies previous approaches in the literature and has already been used, in various guises, in recent research work. The presented methods are related to the layers in the framework and the stages in the development lifecycle of interactive systems. The paper also discusses practical issues surrounding the employment of the above, and provides a brief overview of complementary and alternative approaches in the literature.

Keywords: layered evaluation, evaluation framework, formative evaluation methods, design

1. Introduction

The importance and benefits of involving users in the design and evaluation of adaptive systems has been advocated for a long time (Chin, 2001; Weibelzahl, 2001, 2005; Masthoff, 2002; Gena 2005; Gena & Weibelzahl, 2007). In fact, user studies have become an integral part of papers published in the UMUAI journal, and indeed most papers published in the major conferences in the area. For example, a review of the last three years of UMUAI shows that all papers (excluding surveys and special issue introductions) now contain evaluations, compared to only one third when Chin (2001) surveyed UMUAI for the nine years preceding 2001. Although this is most definitely indicative of increasing maturity in the field, we are far from having solved all related outstanding issues. This paper discusses some of these issues, and proposes a specific evaluation approach and methods for addressing them.

From early on in the history of the field, it has been acknowledged that the evaluation of interactive adaptive systems[1] (IAS) is, in most cases, a complicated endeavour, that is significantly different to the evaluation of non-adaptive interactive systems (see e.g., Totterdell & Boyle, 1990). The differences are attributable to the nature of adaptivity and the implications it has on interaction. In particular, a mainstay of evaluation approaches in Human-Computer Interaction (HCI) is that an interactive system's state and behaviour are only affected by direct and explicit actions of the user. This principle does not hold true in adaptive systems, however. The very aim of adaptivity is to imbue a system with the type of intelligence that allows it to actively take the initiative in supporting the users' activities, on the basis of inferred information about the user and the interaction context, often derived from implicit interaction cues. It is this capacity of adaptive systems to exhibit their own, not directly user-controlled behaviour that traditional evaluation approaches fail to address. Moreover, the adaptation process often takes time, as the system needs to learn about the user's goals, knowledge or preferences, etc., before adaptation can take place. Thus, the observation of any effects of adaptivity may require long-term, or even longitudinal studies, or be based on evaluation designs that explicitly account for that factor. Due largely to these disparities between interactive systems in general and adaptive systems in particular, adaptation was not sufficiently addressed in early standardized evaluation frameworks (although, in some cases, it was a concern) (Stary & Totter, 1997).

To remedy these problems, early approaches to the evaluation of IAS were in the direction of comparative assessments between adaptive and "static" systems. This gave rise to the popular, but potentially also problematic, "with and without" adaptivity evaluation design, in which an adaptive instance of the system is compared with a non-adaptive one. This evaluation design has been used in several studies in the field, including, for example, Kaplan, Fenwick, & Chen (1993), Boyle & Encarnacion (1994), Weber & Specht (1997), and Brusilovsky & Eklund (1998).

A partial summarization of the potential problems that Totterdell & Boyle (1990) associate with comparing adaptive systems with static counterparts, or static instances of themselves, is as follows:

- *Selection of non-adaptive controls*: An adaptive system's behaviour can range over a set of possible states for any given dimension of adaptation. The question, therefore, is which of these states the evaluator should choose for the non-adaptive control. Where

[1] We will be using the term "interactive adaptive systems" throughout this paper to refer to systems that have an interactive front-end and are capable of self-adaptation (applied to, or experienced through, the aforementioned interactive front-end). We further assume that adaptation in such systems is based at least on the characteristics of users (treated individually or collectively), without excluding any other category of adaptation determinants.

appropriate the state might be selected by best current practice. However, there may not always be a plausible control, particularly if the system is a novel application. Furthermore, in all but the simplest situations there will be a very large space of potential system states, which complicates the selection of one of these to serve as the "best" non-adaptive state. Additionally, a non-adaptive instance of a system designed to be adaptive may not be "optimal" in any way, if adaptation is properly designed into the system (Höök, 2000).

- *Selection of equilibrium points*: Another related problem is the selection of appropriate "points of equilibrium" in the evolution of the adaptive system's behaviour to compare against. This often needs to explicitly take into account an initial period of inefficiency during which the system acquires a model of the user (and any other external factors that guide the system's behaviour), and also periods of "flux" during which changes in the user's or system's behaviour have mutual effects that may lead to new points of equilibrium.
- *Dynamics of adaptive behaviour*: Adaptive systems often have to adapt to at least two mutually incompatible criteria (e.g., controllability vs. unobtrusiveness). Thus, enhancements brought about by adaptation and explainable in terms of information about a particular user, group, etc., if applied to another user, group, or task, might instead have detrimental effects. The evaluator then has to show, not only that adaptation is of benefit, but also that there exist different "optima" in the environment, and that the system can find them (e.g., different levels of trade-offs between controllability and unobtrusiveness that would be "optimum" for a given user or category of users). Combining this requirement with the fact that adaptive behaviour evolves over time, there is a multiplicative effect on the number of states in which the system is, as Totterdell & Boyle (1990) term it, "compatibly adaptive" to its environment; in a comparative assessment approach, all these states would ideally be targeted by evaluation.

Further to the above, an implicit assumption of the comparative assessment approach is that a system "converges" to a state that can then be compared. This, however, leaves other desirable attributes of adaptation unaccounted for, such as the system's capacity to detect changes in its environment, and smoothly transition to new states of convergence, neither exhibiting oversensitivity to minor fluctuations, nor reacting so slowly as to cause long periods of mismatch between its behaviour and its environment.

These problems may be difficult to address in certain IAS domains, but do not, in fact, render the employment of comparative approaches in studies prohibitive. Although, to the best of our knowledge, such work has not as yet been reported in the literature, approaches that would allow for a systematic selection of states to include in comparisons would be within reach of the research community.

One point that requires further attention is that, when a comparative approach is employed, then, by definition, the question asked is a variation of: "is this (adaptive) version better than that (non-adaptive) version, in this particular respect?" This is indeed a fundamental question and a defining one in establishing the "value" of adaptation in particular settings. However, it may not provide sufficient insights in terms of the fine-grained effects of adaptive system behaviour, and the findings may not be readily generalisable beyond the specific adaptation settings and behaviour of a single system. More specifically, when employing comparative assessment without directly addressing specific aspects of adaptation, the reasons behind the "success", or "failure" of adaptation can only be traced back to the initial hypotheses of the adaptive system design. In other words, it may not be possible to ascertain why, and under what conditions, a particular type of adaptation may be employed towards a specific goal. This constraint may be prohibitive in cases where evaluation is intended to derive design knowledge that can be fed back into the system's development

process. In short, then, we can say that comparative assessment can be potentially very useful, probably even more so in the context of evaluating the system against the overall goal that adaptation was introduced to achieve. However, when it is applied at that level, it may not be able to offer the type of insight necessary for attaining and validating adaptation design knowledge.

A major characteristic of evaluations that was alluded to above is their goal. A widely accepted coarse classification uses an evaluation's goal to distinguish between *formative* and *summative* evaluations (Scriven, 1996). Formative evaluation aims to identify shortcomings or errors in a system in order to further improve it and to guide the system design and development. In contrast, summative evaluation aims to determine the value or impact of a system. Formative evaluation goes hand-in-hand with the HCI principle of involving users as early as possible in the design process, and is vital in discovering what and how to improve in an interactive system (Gould & Lewis, 1985; Shneiderman, 1998). Whereas summative evaluation is well established and in wide use, the same is not true of formative evaluation. Most user-based assessments of IAS in the literature report only summative evaluations, aiming to establish the extent to which the use of an adaptation method has improved the system, or the extent to which the user modelling is accurate. Some recent notable exceptions of papers that include formative studies that appeared in UMUAI include (Stock et al., 2007), (Carmagnola et al., 2008) and (Porayska-Pomsta, Mavrikis & Pain, 2008), while some others mention that a formative study has preceded the summative one, but do not report its results (e.g., Kosba, Dimitrova & Boyle, 2007).

Although the inherent difficulties in the evaluation of adaptation, as discussed thus far, have been well understood for quite some time, no satisfactory solutions or principled alternative approaches emerged until the beginning of the last decade. During that time, empirical studies that evaluated IAS remained few, and, more often than not, provided ambiguous results. In the last ten years, the evaluation of IAS started receiving considerable renewed attention. This has been due, in part, to the increasing utilization of adaptivity methods and techniques in a wide range of application domains, but also due to the desire to acquire a solid design basis for adaptation, unattainable until the largely unsolved problems involved were addressed (see, e.g., Brusilovsky & Eklund, 1998; Höök, 2000; Chin, 2001; Masthoff, 2002).

This last decade has seen the introduction of a number of promising attempts at tackling the problem of evaluating IAS, sharing one main idea: to treat adaptation not as a singular, opaque process, but, rather, "break it down" into its constituents and evaluate each of these constituents separately where necessary and feasible. These approaches became known under the moniker of "layered" evaluation of adaptive systems.

An oft-cited example of the application of the related principles and their potential benefits are two studies of the same system, one following a layered evaluation approach and one not. The first study on the effects of adaptive link annotation (described in Brusilovsky & Eklund, 1998) demonstrates well the problems that can arise when evaluating an adaptive system. This study treated the adaptation process as a "monolithic" entity and aimed to assess it as a whole. Specifically, the goal of that experiment was to assess the impact of (link-oriented) adaptive navigation support (ANS) on students' learning and on their paths through the learning space. Contrary to expectations, the study failed to show any statistically significant differences between the versions with and without ANS. Although the authors did perform additional analysis and offered some potential justifications for their findings, the matter remained largely inconclusive. A revisited interpretation of the initial study was then presented (Brusilovsky, Karagiannidis, & Sampson, 2001), which decomposed the adaptation

process into two layers that were evaluated separately. This study demonstrated that whereas the user models created were sufficiently accurate, the adaptations applied on the basis of these models were likely not appropriate for the target population.

The above and other propositions on how layered evaluation of IAS can be approached have been directly or indirectly in use for some time now. This paper attempts to unify and organize the principles of layered evaluation, as these emerge from the different propositions and related work in the literature, into a framework that is based on a decomposition model of the adaptation process that identifies five stages or layers in the process. It also presents an array of evaluation methods that can be used in association with the proposed framework.

More specifically, the rest of the paper is structured as follows. We start by outlining the history of layered evaluation, and the underpinnings of the specific framework presented herein (Section 2). Following that, we present the proposed framework, providing a rationale and a basis for the evaluation of each of the identified layers, and propose a number of generic criteria that can be evaluated in relation to each layer (Section 3). We then provide an extensive overview of evaluation methods that can be tailored, or have been specifically developed to cater for the idiosyncrasies of evaluating adaptive systems; we focus on methods suitable for formative evaluation, and relate these to the proposed framework's layers, and the stages in the development lifecycle of interactive systems (Section 4). We then address practical issues related to the employment of the framework, including the derivation of application domain- and adaptation type- specific criteria, the tailoring of the layered approach to suit individual evaluation requirements, and the selection of appropriate evaluation methods for different layers and development stages (Section 5). We next turn our attention to limitations of the framework and the general evaluation approach put forward in this paper, including the evaluation methods presented, and list some of the complementary and alternative approaches that can be used to address these shortcomings (Section 6). Finally, we discuss the impact of layered evaluation in the literature thus far, potential benefits of its application, and related work in the literature (Section 7).

2. History and Underpinnings of Layered Evaluation

The seeds of the idea of decomposing adaptation for evaluation purposes can be traced back to Totterdell & Boyle (1990), who propose that a number of adaptation metrics be related to different components of a logical model of adaptive user interfaces, to provide what amounts to adaptation-oriented design feedback. Furthermore, Totterdell & Boyle (1990) present two types of assessment performed to validate what is termed "success of the user model" (note that, in their case, the "user model" is also responsible for adaptation decision making): "Two types of assessment were made of the user model: an assessment of the accuracy of the model's inferences about user difficulties; and an assessment of the effectiveness of the changes made at the interface." (Totterdell & Boyle, 1990, p. 180)

This main idea remained dormant for several years, but was revived and further pursued in the past decade, in an attempt to resolve the problems encountered when employing methods and techniques intended for "traditional" interactive systems to their adaptive counterparts.

As already mentioned, Brusilovsky et al. (2001, p.3) advocated layered evaluation, "where the success of adaptation is decomposed into, and evaluated at, different layers, reflecting the main phases of adaptation […]" (see **Error! Reference source not found.**). The authors describe the identified layers thusly (Brusilovsky et al., 2001) (emphasis by the authors):

- In the *interaction assessment layer*, only the assessment phase is being evaluated. That is, the question here can be stated as: "are the conclusions drawn by the system concerning the characteristics of the user-computer interaction valid?" or "are the user's characteristics being successfully detected by the system and stored in the user model?"

– In the *adaptation decision making layer*, only the adaptation decision making is being evaluated. That is, the question here can be stated as: "are the adaptation decisions valid and meaningful, for selected assessment results?"

Simultaneously with the aforementioned idea, two related evaluation frameworks were proposed. The first was a process-based framework presented by Weibelzahl (2001), which discerned four layers that refer to the information processing steps within the adaptation process (**Error! Reference source not found.** – note that in this figure the steps are represented by arrows, whereas, in the rest of the figures in this section, they are represented by rectangular "nodes"):

– *Evaluation of input data* (Step 1 in **Error! Reference source not found.**), refers to the evaluation of the reliability and external validity of the input data acquisition process, as well as of the acquired data itself.
– *Evaluation of the inference mechanism* (Step 2 in **Error! Reference source not found.**), addresses the evaluation of the validity of user properties inferred from the input data previously collected.
– *Evaluation of the adaptation decisions* (Step 3 in **Error! Reference source not found.**), deals with determining whether adaptation decisions made are optimal, determined through the comparison of possible alternative decisions based on the same specific set of inferred user properties.
– *Evaluation of the total interaction* (Step 4 in **Error! Reference source not found.**), finally, is geared towards the summative assessment of adaptation, and distinguishes between the evaluation of system behaviour (including factors such as the frequency of adaptation), and the evaluation of user behaviour (as affected by adaptation) and the system's overall usability.

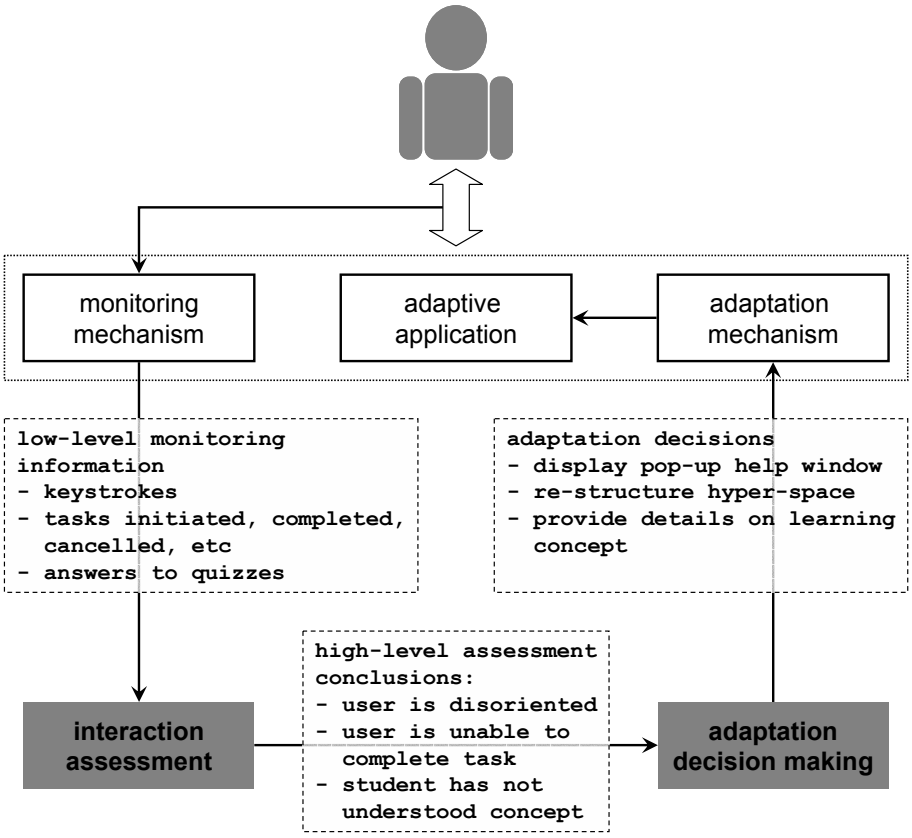

Figure 1: Adaptation decomposed into two high-level phases: interaction assessment and adaptation decision making (Brusilovsky et al., 2001)

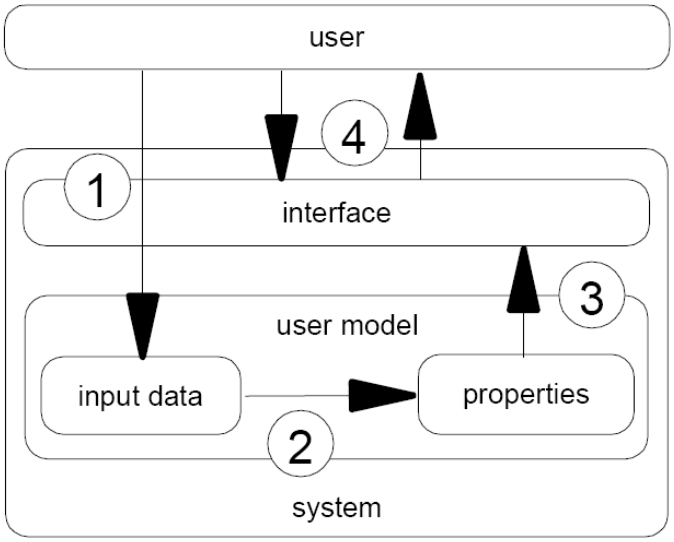

Figure 2: Four-layered decomposition model in the evaluation framework proposed by Weibelzahl (2003).

This framework has a very clear focus on the empirical evaluation of IAS and has been applied in practice to different adaptive learning courses, including several studies with thousands of users (Weibelzahl & Weber, 2003).

The second framework proposed around the same time by Paramythis, Totter and Stephanidis (2001) adopts a more engineering-oriented perspective in the identification of layers (termed "modules" in the respective paper), focusing in more detail on the different components involved in the adaptation process (**Error! Reference source not found.**). The framework identifies the following stages/components of adaptation in adaptive user interfaces:

- *Interaction monitoring*, encapsulates the collection of input data.
- *Interpretation/inferences*, refers to inferences drawn upon the collected input data.
- *Modelling*, refers to the population of user-, context- and other dynamic models, as well as to the utilization of any static models (e.g., a domain- or task- model)
- *Adaptation decision making*, captures the process of making high-level adaptation decisions (e.g., identify products that are likely of interest to the user), on the basis of the available models.
- *Applying adaptations*, refers to "instantiating" adaptation decisions into the system (e.g., showing a panel with the list of recommended products, or promoting them in a list including other products).

Based on these, the framework then goes on to suggest evaluation "modules" that address the evaluation of these adaptation stages in isolation or in combination. This framework also discusses the issue of formative vs. summative evaluation, and makes some initial suggestions as to which (of the then existing) methods and tools might be appropriate for the evaluation of different adaptation modules, in order to elicit input for the development process.

The frameworks discussed thus far have several significant differences, both in the stages of the adaptation process they seek to highlight and address, and in the evaluation approaches they (implicitly or explicitly) advocate. However, there is inarguably also a lot of common ground: the premise of all these frameworks is that adaptation needs to be decomposed, so that its comprising stages/elements can be assessed/evaluated in isolation. Their main

conceptual differences lie with the decomposition models used, and, in particular, with the models' perspectives on adaptation, as well as with the adopted level of granularity. **Error! Reference source not found.** provides a pictorial representation of the differences and relations between the decompositions proposed by these three frameworks.

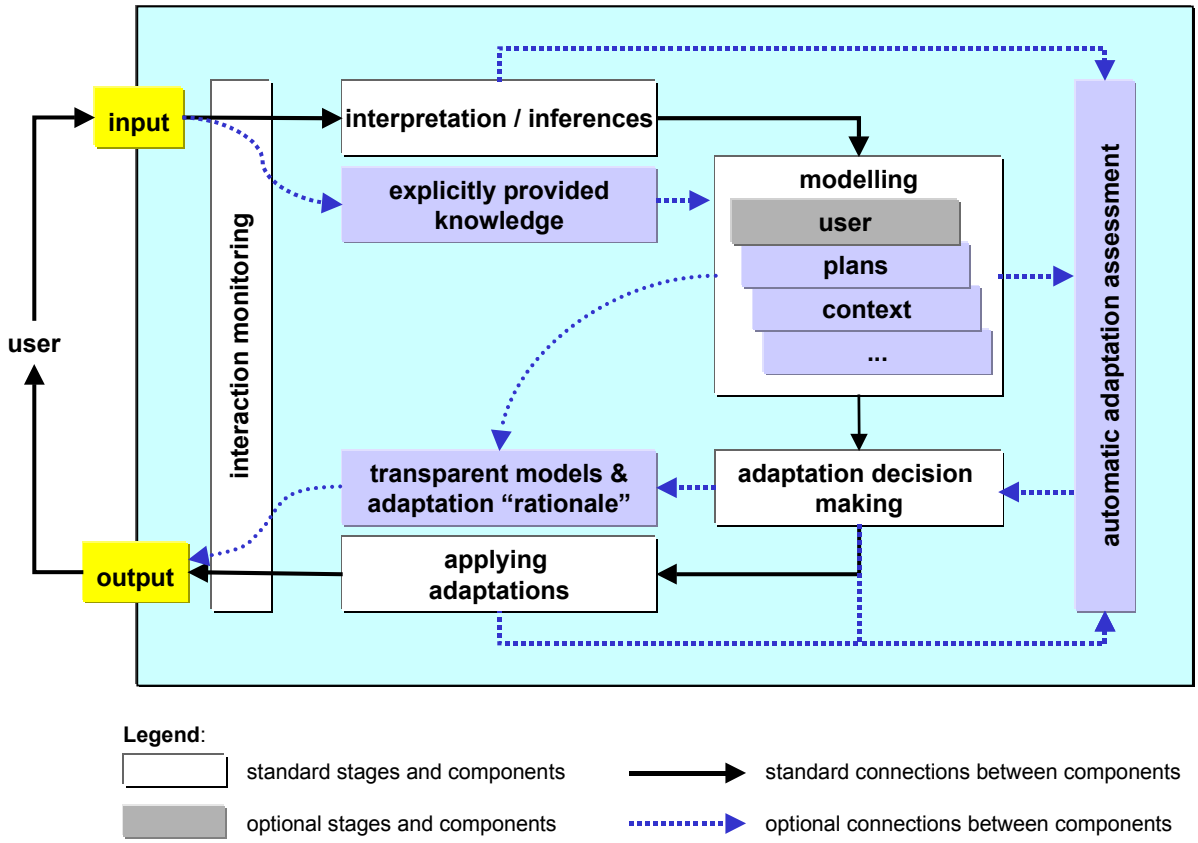

Figure 3: Decomposition model for "modular" evaluation of adaptive user interfaces (Paramythis et al., 2001)

Paramythis & Weibelzahl (2005) presented the first steps of an effort to merge or unify the common themes of these frameworks. These efforts towards a unification of the alternative propositions, culminating into the framework proposed in this paper, are based on the introduction of a model of decomposition with the widest possible applicability on existing and forthcoming IAS, making few assumptions about implementation and architectural properties of the system, but, at the same time, offering a concrete enough guide to evaluation activities.

To arrive at the desired decomposition model, we have examined not only the previously proposed frameworks, but also the common properties of existing models and architectures for adaptation. Although relatively young, the field of IAS is abundant with conceptual, architectural, and functional models of adaptation, spanning a large range of theoretical approaches to adaptation, types of adaptation supported, component technologies, etc. (see, e.g., Totterdell & Rautenbach, 1990; Oppermann, 1994; De Bra, Houben, & Wu, 1999; Jameson, 2001; Koch & Wirsing, 2002; Knutov, De Bra, & Pechenizkiy, 2009). This

pluralism is further compounded by the existence of domain- and "platform"[2]- specific models/architectures, which cannot be easily generalised or extended in their coverage. For example, several reference models have been developed for adaptive hypermedia (e.g., De Bra et al., 1999; Koch & Wirsing, 2002; Ohene-Djan, 2002), but currently these are not generally applicable to adaptive systems and are rather intended to support software engineers in developing systems.

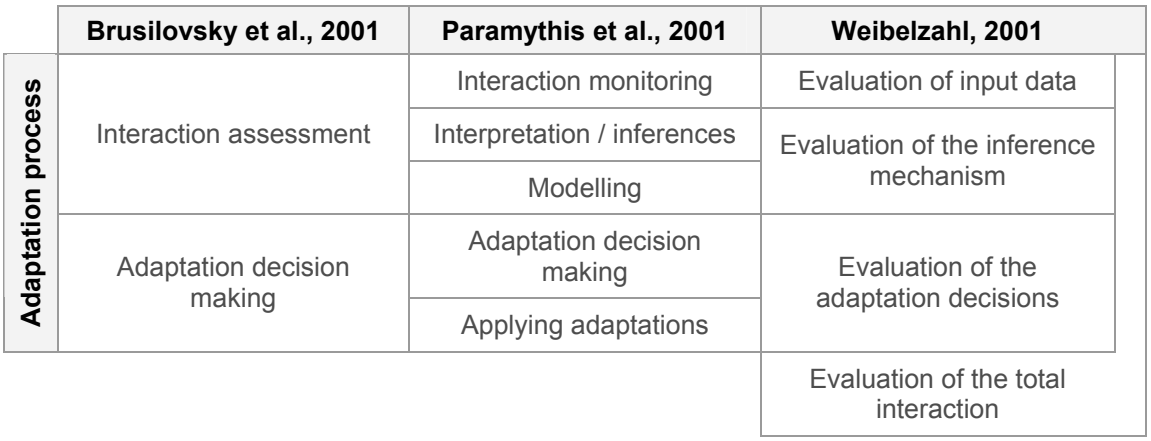

Figure 4: Comparison of the adaptation decomposition models in the three frameworks presented by Brusilovsky et al. (2001), Paramythis et al. (2001) and Weibelzahl (2001).

In examining adaptation models in the literature, we have restricted ourselves to the process-oriented ones (as opposed, for instance, to component-oriented ones), so as to allow for the maximum possible degree of flexibility in terms of how adaptation is implemented (where, in fact, approaches proliferate). We concentrate here on three models (or architectures) of adaptive systems that have been proposed in the literature:

- A very early process-oriented architecture was put forward by Totterdell and Rautenbach (1990) (**Error! Reference source not found.**), which was the basis for the framework proposed by Paramythis et al. (2001) (see also **Error! Reference source not found.**). This model relates major architectural elements of adaptive user interfaces in a multi-step adaptation process involving the collection of input data (*Interaction cues*), the creation/utilization of models (*User/Task Models*), and the selection of appropriate adaptive interventions (*User Interface Variants*), all on the basis of the underlying *Adaptive Theory*.
- Another proposal by Oppermann (1995) describes adaptive systems as consisting of three parts: an *afferential*, an *inferential* and an *efferential* component. According to (Oppermann, 1995, p. 6), "[t]his nomenclature borrows from a simple physiological model of an organism with an afferential subsystem of sensors and nerves for internal and external stimuli, with an inferential subsystem of processors to interpret the incoming information, and with an efferential subsystem of effectors and muscles to respond to the stimuli".
- More recently, Jameson (2001) presented a "general schema" for the processing in user-adaptive systems (**Error! Reference source not found.**), which can be informally described as follows (Jameson, 2008, p. 433): "A user-adaptive system makes use of some type of information about the current individual user, such as the products that the user

[2] The term "platform" is used here in its general sense. Exemplifying this use, we would categorise, for instance, the "Web" as one such platform, quite distinctly from the "desktop" platform.

has bought. In the process of user model acquisition, the system performs some type of learning and/or inference on the basis of the information about the user in order to arrive at some sort of user model, which in general concerns only limited aspects of the user (such as her interest in particular types of product). In the process of user model application, the system applies the user model to the relevant features of the current situation in order to determine how to adapt its behaviour to the user."

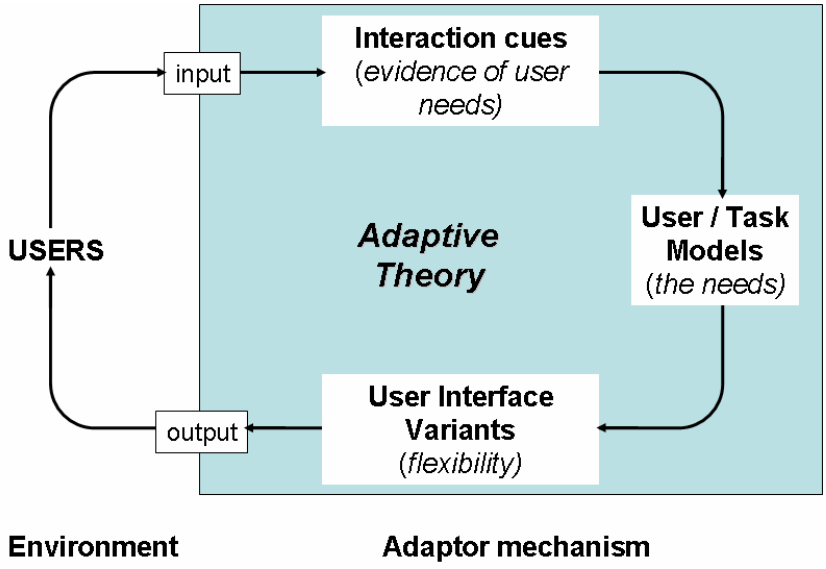

Figure 5: Logical two-level architecture of adaptation;
adapted from (Totterdell, Rautenbach, Wilkinson, & Anderson, 1990)

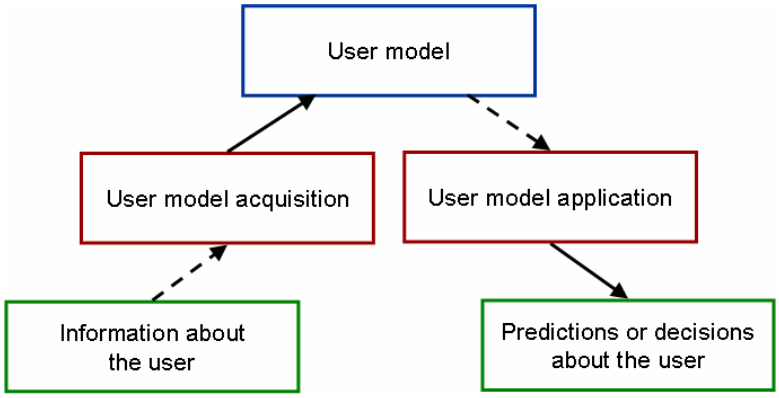

Figure 6: General schema for the processing in a user-adaptive system; adapted from (Jameson, 2008).
(dotted arrows: use of information; solid arrows: production of results.)

These models represent different points of view, and focus on different aspects of adaptation. One important similarity that they do have, though, is that they do not attempt to be prescriptive in terms of the modules/components that make up an adaptive system. Instead, they focus, directly or indirectly, on the "steps" or stages of the adaptation process in interactive adaptive systems.

Even more importantly, the models under discussion exhibit a number of common characteristics:

- They commence with the collection and interpretation of "observation data", which, in these models, relate mainly to the user's behaviour (see "Interaction Cues" in **Error! Reference source not found.**, and "Information about the user" in **Error! Reference source not found.**).
- In all three cases, there is an "inference" step, which results in the creation or updating of corresponding models, on the basis of the observed data (see "User/Task Models" in **Error! Reference source not found.**, and "User model acquisition" in **Error! Reference source not found.**). Typically, this involves the employment of an intelligent mechanism that infers user-, context-, etc., characteristics from the raw data.
- Split between Oppermann's (1995) "inferential" and "efferential" steps, and represented individually in the other two models (see "User Interface Variants" in **Error! Reference source not found.**, and "User model application" in **Error! Reference source not found.**), is the task of making decisions as to how the system should be adapted, i.e., how the system behaviour should be changed.

The identified common characteristics of the above models, coupled with the precursor work on layered evaluation frameworks for IAS, form a solid basis for the proposal described in detail in the next section.

3. The Proposed Evaluation Framework

3.1. A Model for "Decomposing" Adaptation

As already discussed, a comprehensive, yet not prescriptive, model of adaptation is of paramount importance to the framework at hand. We have composed this model by factoring out and enriching the common elements of previous attempts and the related models outlined in the previous sections. Its foundations lie on the identification of three rough categories of "activities" in the adaptation process in an IAS: observing and interpreting (user) input; adjusting internal models that evolve on the basis of that input; and, using the up-to-date models to determine the system's adaptive behaviour. This rough set has been elaborated upon and refined to better capture elements of the adaptation process that may need to be assessed. The resulting model is depicted in **Error! Reference source not found.**. Briefly, the main layers of adaptation identified are (**Error! Reference source not found.**):

(a) *Collection of input data* (CID) refers to the assembly of user interaction data, along with any other data (available, e.g., through non-interactive sensors) relating to the interaction context.
(b) *Interpretation of the collected data* (ID) is the step in which the raw input data previously collected acquire meaning for the system.
(c) *Modelling of the current state of the "world*" (MW) refers to derivation of new knowledge about the user, the interaction context, etc., as well as the subsequent introduction of that knowledge in the "dynamic" models of the IAS.
(d) *Deciding upon adaptation* (DA) is the step in which the IAS decides upon the necessity of, as well as the required type of, adaptations, given a particular state of the "world", as expressed in the various models maintained by the system.
(e) Finally, *applying* (or *instantiating*) *adaptation* (AA) refers to the actual introduction of adaptations in the user-system interaction, on the basis of the related decisions.

It is argued that each of these adaptation layers needs to be evaluated explicitly, although not all layers can be "isolated" and evaluated separately in all systems. Furthermore, the nature of the IAS will necessarily dictate the relevance of each of these layers.

Before we move on with the discussion of each of the layers, it is important to make some preliminary remarks on the rest of the elements that appear in the model. Firstly, the figure contains several elements "internal" to the IAS ("static" and "dynamic" models, and adaptation theory). These are briefly described below.

The models potentially maintained by the IAS are separated into two broad categories. The first category groups together the IAS's "static" models (comprising, for instance, the system model, the task model, the application model, etc.) These are often implicit, i.e., there does not necessarily exist an explicit representation of them in the IAS; rather, they may be "dispersed" in the form of domain knowledge throughout the system. In several cases, of course, explicit representations do exist and are actively used in deciding upon adaptations (e.g., in the case of user plan recognition, a task model is a necessity). This first category of internal IAS models is used, again implicitly or explicitly, when interpreting input data. Consider as an example the case of an adaptive, Web-based course delivery system; the fact that the user has requested a specific URL may be interpreted by the system as a request for viewing/reading the contents of the corresponding organization of learning material(s).

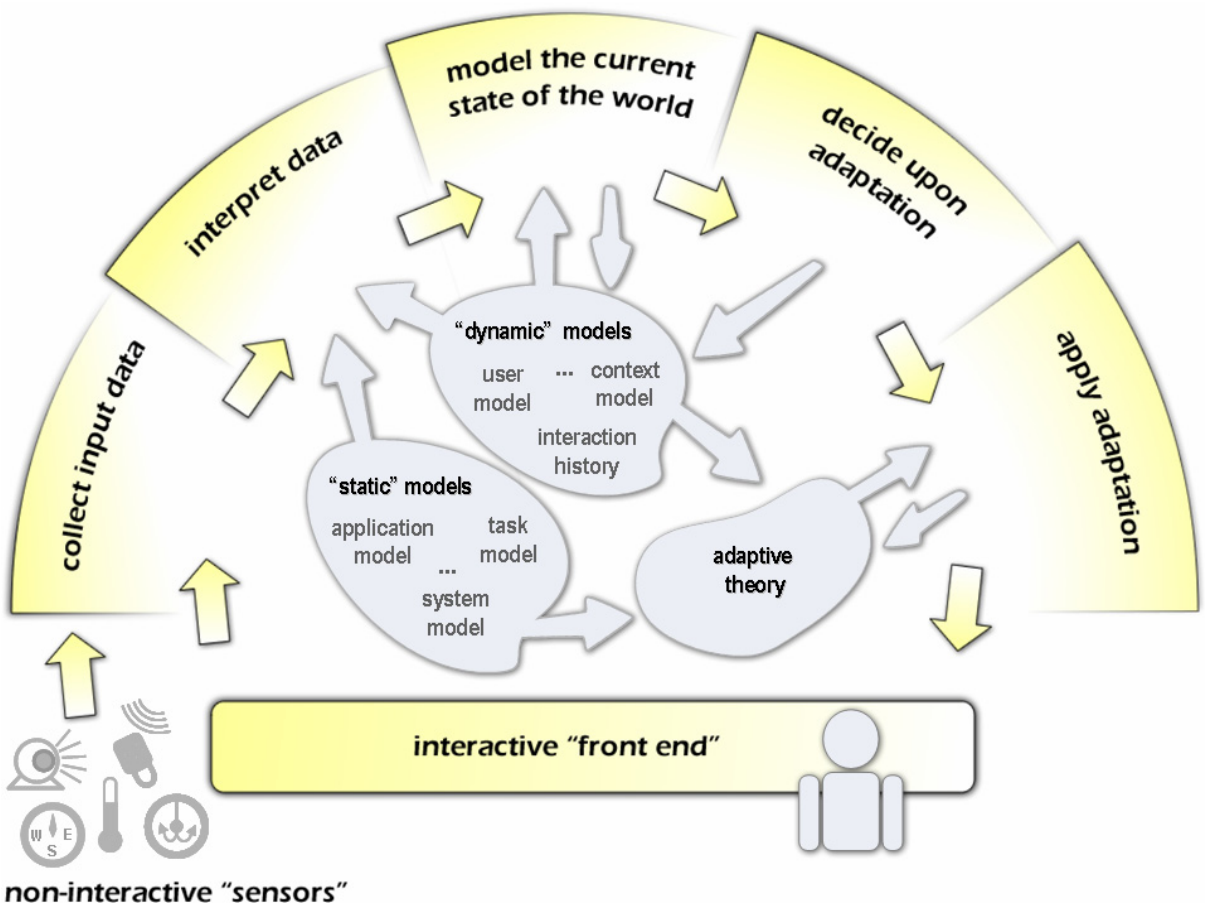

Figure 7: The adaptation decomposition model underlying the proposed evaluation framework.

The second category groups together the IAS's "dynamic" models (comprising, for instance, the user model[3], the context model[4], a representation of the interaction history, etc.) These are models that are updated by the IAS, on the basis of new knowledge that it derives from the interpretation of the input data. They are, typically, the main determinant for adaptation decisions, and can be used in various ways in the decision-making process (for example, a user model can be used to decide upon adaptations for a specific user, or be combined with models of other users to provide support for decisions based on the characteristics, behaviour, etc., of entire groups of users).

Error! Reference source not found. also introduces an entity termed "adaptive theory". The term itself is borrowed from (Totterdell & Rautenbach, 1990) and is used to refer to the theory that underlies adaptations in the system (see also **Error! Reference source not found.**). The word theory is not used here in its formal sense, but rather to represent the totality of adaptation goals/objectives that drive adaptation in the IAS. In several systems, the adaptive theory is dispersed into possibly independent adaptation "rules" which are themselves "triggered" by the contents of the IAS's models (e.g., the user model).

Finally, arrows are used in the figure to denote potential flows of information. Although some of the depicted flows will be typical in certain categories of IAS, or in certain application domains, only part of them are usually present in any one system. For example, the flow from the adaptation decision layer, to the "adaptive theory" entity, exists only in IAS that have a, so-called, second adaptation cycle (Totterdell & Rautenbach, 1990) - i.e., systems which are capable of assessing their own adaptation decisions and modifying their adaptation strategies.

Note that *the above described elements are not part of the model itself*. Their inclusion in the figure is solely intended to facilitate understanding of the model and support related discussions. The proposed decomposition model (and, consequently, the proposed framework) is neither based upon, nor presupposes the presence of any of the models identified in the figure (with the possible exception of the user model, or its equivalent). Further, we explicitly do not assume specific approaches to intelligence, or decision making, although the depiction of the model might suggest that. In fact, different approaches along the above lines might lead to different groupings of the layers, which, for instance, may happen collectively, or have but rudimentary manifestations in an IAS. The subsequent discussion of the adaptation decomposition model is explicitly based on these provisions.

3.2. Layered Evaluation of Adaptation

In this section we will present in more detail each of the layers that appear in the model and discuss whether they need to be evaluated (in isolation or combination) and with what objectives. To this end, we will also introduce specific evaluation criteria that can potentially serve as "guides" for their respective evaluation "layers". Criteria that we believe are applicable to all layers are discussed separately, after the layers themselves. Discussions concerning assessment methods that might be appropriate for evaluating the proposed criteria, as well as the scope and practical use of the framework and its relation to the specific application domain of the IAS, are deferred until later sections. **Error! Reference source not**

[3] It should be noted that, in some categories of adaptive systems, the user model is created once and does not evolve over time. In these cases, one might categorize the user model with the static models of the system. Nevertheless, these models can still be treated as dynamic, since they refer to individual users and are not "shared" among users (as is the case, for instance, with a system's task model).

[4] The term "interaction context" (often shortened to "context" in this paper) is used to refer to all information relating to an interactive episode that is not directly related to an individual user. This interpretation of the term follows the definition of context given by Dey & Abowd (2000), but diverges in that users engaged in direct interaction with a system are considered (and modelled) separately from other dimensions of context. Therefore, the interaction context is characterised, for example, by: features and capabilities of access terminals, characteristics of network connections, the user's current location, current date/time, etc.

found. is intended to act as a "guide" to the rest of this section, and provides a collective overview of the layers and the proposed criteria for each of them, along with the formative evaluation methods that may be applicable in each case. To facilitate reading, the relevant portions of the table are repeated at the end of the discussion of each layer.

3.2.1. Collection of Input Data

The "input" data that an interactive system collects is predominantly derived from the user's interaction with it, i.e., it comes from direct interaction of the user with the system's user interface, or interactive front end[5]. Data in this category include the user's pression of a button, selection of a link, etc. It is important to note at this point that input data of this nature does not necessarily carry any semantic information. It is in the next layer, and with the assistance of (implicit or explicit) application- and task- models that this low-level data will acquire "meaning" for the system.

In addition to the traditional input data that an IAS may derive from direct user interaction, there exists a host of additional information that may be available to the system, from "sensors" not directly or explicitly manipulated by the user. For example, in an adaptive environment, the user's position, direction of movement, gestures, direction of gaze, etc. may also be available (Zimmermann & Lorenz, 2008); smart environments such as smart homes or smart offices often rely on a variety of sensors. The *accuracy* of these sensors needs to be considered, before the resulting data can be used for further inferences. For example, the thermometer measuring the office temperature in an intelligent office environment had an error of about 2°C (Cheverst et al., 2005) and the positioning system in an adaptive museum guide provided a resolution of 5cm and 5° in terms of orientation (Zimmermann, Specht, & Lorenz, 2005). In fact, the quality of the input data can be evaluated in a systematic way. Using the example of an adaptive museum guide, Schmidt, Zukerman, & Albrecht (2009) describe how the impact of uncertainty in sensing technology in physical spaces can be investigated. With more and more sensor data available for user modelling, a number of IAS have been described that take advantage of such potentially useful information as the ambient noise in the user's environment (Cheverst et al., 2005), the presence of the user in front of the interaction "terminal" (Oliver & Horvitz, 2005), the very fact that the user is looking at the screen or not, and even the user's affective state inferred from physiological sensors (Cooper et al., 2009). In many cases, the accuracy of these sensors seems to be taken for granted. In fact there may be no need to evaluate this layer if previous studies have shown that the data is reliable or if it is safe to assume that the data is reliable. Presence of data in this category may not always directly affect user modelling itself, but most certainly does affect the interpretation of user-related data, or may even be used to model the broader context of interaction.

Table 1: An overview of layers and related criteria, along with methods that can be used for their evaluation.

Layer	Goal	Evaluation criteria	Evaluation methods

[5] In the rest of this paper we will be using the term "interactive front end", rather than "user interface". This is done to explicitly denote a potentially richer interactive experience than the one afforded by today's WIMP, keyboard- and mouse- based user interfaces, as well as to avoid misinterpretations that may result from the typical association of the term "user interface" to desktop-based interaction.

Collection of Input Data (CID)	Check quality of raw input data	Accuracy, latency, sampling rate	Data Mining (see 4.1.3); Play with Layer (see 4.3.1); Simulated Users (see 4.3.2); Cross-Validation (see 4.3.3)
Interpretation of the Collected Data (ID)	Check that input data is interpreted correctly	Validity of interpretations, predictability, scrutability	Data Mining (see 4.1.3); Heuristic Evaluation (see 4.2.2); Play with Layer (see 4.3.1); Simulated Users (see 4.3.2); Cross Validation (see 4.3.3)
Modelling the Current State of the "World" (MW)	Check that constructed models represent real world	Primary Criteria: Validity of interpretations or inferences, scrutability, predictability; Secondary Criteria: Conciseness, comprehensiveness, precision, sensitivity	Focus Group (see 4.1.1; 4.2.1); User-as-Wizard (see 4.1.2); Data Mining (see 4.1.3); Heuristic Evaluation (see 4.2.2); Play with Layer (see 4.3.1); Simulated Users (see 4.3.2); Cross-Validation (see 4.3.3)
Deciding upon Adaptation (DA)	Determine whether the adaptation decisions made are the optimal ones	Necessity of adaptation, appropriateness of adaptation, subjective acceptance of adaptation, predictability, scrutability, breadth of experience	Focus Group (see 4.1.1; 4.2.1); User-as-Wizard (see 4.1.2); Heuristic Evaluation (see 4.2.2); Cognitive Walkthrough (4.2.3); Simulated Users (see 4.3.2); Play with Layer (see 4.3.1); User Test (see 4.3.1)
Applying Adaptation Decisions (AA)	Determine whether the implementation of the adaptation decisions made is optimal	Usability criteria, timeliness, unobtrusiveness, controllability, acceptance by user, predictability, breadth of experience	Focus Group (see 4.1.1); User-as-Wizard (see 4.1.2); Heuristic Evaluation (see 4.2.2); Cognitive Walkthrough (4.2.3); User Test (see 4.3.1); Play with Layer (see 4.3.1)
Evaluating Adaptation as a Whole	Evaluate the overall adaptation theory, may be either formative or summative	Specific for system's objectives or underlying theory	Heuristic Evaluation (see 4.2.2); Cognitive Walkthrough (see 4.2.3); User Test (see 4.3.1); Play with Layer (see 4.3.1)
All layers	---	Privacy, transparency, controllability	Focus Group (see 4.1.1; 4.2.1); Cognitive Walkthrough (see 4.2.3); Heuristic Evaluation (see 4.2.2); User Test (see 4.3.1)

The nature of the sensors and the way in which their input is used will typically determine what other criteria may need to be assessed. For instance, excessive *latency* in a GPS sensor may result in the system adapting to a geographical context that "lags behind" the user's current one, and a low *sampling rate* for an accelerometer may have adverse effects in a mobile guide that adapts its output to whether its user is on the move or stationary. It should also be noted that certain categories of sensors, especially ones not normally employed in interactive situations (e.g., ones related to a person's physical well being), may well pose sensor- and context- specific considerations and may require the introduction of respective (possibly entirely custom) criteria.

It is noteworthy that undetected problems in this layer may have "cascading" effects in other layers. Returning to a previous example, treating a user's position in a physical space, as relayed by sensors, as entirely accurate, may lead to problematic interpretations of the users' interests in relation to objects within that physical space. In contrast, when the level of

inaccuracy that should be anticipated is known, it can well be integrated into the adaptation models of subsequent layers, as demonstrated by Schmidt, Zukerman, & Albrecht (2009). These "cascading" effects can arguably occur between most pairs of subsequent layers, but are most often neglected in this layer.

In synthesis, either of the categories of data discussed (i.e., originating from the user, or from non-interactive sensors) is subject to "technical" assessments which would determine whether factors such as *accuracy*, *latency*, *sampling rate*, etc. are appropriate for the system at hand. Given the assumption that "raw" input data does not carry semantic value by itself, such assessments may be all that is necessary at this level. A summary of this layer is given in **Error! Reference source not found.**.

Table 2: Summary of Collection of Input Data Layer

Collection of Input Data (CID)	
Goal	Check quality of raw input data
Evaluation criteria	Accuracy, latency, sampling rate, etc.
Evaluation methods	Data Mining (see 4.1.3); Play with Layer (see 4.3.1); Simulated Users (see 4.3.2); Cross-Validation (see 4.3.3)

3.2.2. Interpretation of the Collected Data

What is far more interesting and challenging in terms of evaluation is the layer of interpretation of the input data. According to the proposed model, this is the very layer at which input data acquire "meaning" of relevance to the system. It should be noted that the distinction between this stage and the collection of the input data may seem somewhat artificial as far as current practice is concerned. It is usually the case that input data is retrieved and interpreted in one step. The separation here is not intended as a proposal for a new engineering paradigm or implementation approach; rather, it seeks to explicitly identify and conceptually dissociate the two stages, thus making it possible to discuss them in isolation.

The interpretation process may be straightforward, in those cases that there exists a direct, one-to-one mapping between the raw input data and their semantically meaningful counterparts. Examples include the retrieval of a user's position (when the latter is regarded in its strict geographical confines), the identification of a user action in the context of a task, etc. When the interpretation is unambiguous, and independently of whether it employs any of the system's "static" models, it can be assessed objectively and in a user-independent manner. For instance, an adaptive user support system (Encarnação & Stoev, 1999) might exploit action sequences; registering the number of sessions that the user completed or the number of the user's undo actions is probably highly reliable. There is no subjective judgment or other noise involved in this observation.

Potential problems arise when: (a) the interpretation makes use of assumptions, or (b) the interpretation requires some level of inference. Assumptions and inferences are quite commonly employed in existing IAS, mainly due to the lack of additional data that can better describe the context of interaction. A typical example is how adaptive Web-based information systems consider a node in the hypermedia space "visited", "learned", "of (no) interest", etc., on the basis of how long the user spent on viewing the respective page. Another example is sensor data in intelligent homes, which is particularly difficult to interpret (Sixsmith, 2000).

Although considerable work has gone into developing and proving principles upon which "educated" assumptions or inferences can be drawn (see, e.g., Goecks & Shavlik, 2000; Claypool, Le, Wased, & Brown, 2001; Spada, Sánchez-Montañés, Paredes, & Carro, 2008), these are always dependant on the IAS's application domain, deployment context, etc.

A criterion that may need to be addressed at this stage is the *validity of the interpretations* (at least in cases where the interpretations are not straightforward, as discussed above). An example where validity plays a central role would be an adaptive news broker (Billsus & Pazzani, 1999). Users might provide feedback about a specific news story by selecting one of four categories: *interesting*, *not interesting*, *I already know this*, and *tell me more about this*. But the user's answer depends on many uncontrollable factors. Users might read the story only roughly and might overlook some interesting new facts. Or they might read the same story somewhere else afterwards. Or, just for the moment, they might not be interested in this kind of stories. Several other threats to validity do arise here, and further inferences might be highly biased if the data quality is neither assured nor considered in this process.

Revisiting the example of the "state" of pages that have been viewed by a user, we could now further identify it as a problem of validity of the interpreted data. If the system's assumption (or inference) that a page's content is "known" by the user is erroneous, this could lead to entirely unexpected and unacceptable adaptations, although the rest of the adaptation cycle may be flawless.

Some systems attempt to compensate for potentially erroneous interpretations by explicitly incorporating in them the concept of uncertainty. For example, an IAS might assign interpretations a probability, which might even be related to similar/related interpretations made in the past. Carmichael, Kay, & Kummerfeld (2005) show how the inconsistency of sensor data can be modelled. Modelling the uncertainty can help to identify distortions in sensor data (Schmidt et al., 2009). It is claimed that the proposed evaluation criteria for this stage of adaptation are valid in this case as well, although their scope may need to be adjusted.

When considering the IAS behaviour from the user's perspective, we can identify two additional criteria that may need to be addressed at this layer, namely *predictability of the system's interpretations* and *scrutability of the system's interpretations*. Jameson (2008) defines predictability in this context to represent the extent to which users can predict the effects of their actions. We specialize the definition for the needs of applying this criterion to this layer, and constrain it specifically to the system's interpretation of user actions. When users have a wrong mental model of the principles of this interpretation, there is a very real danger that they will try to modify their behaviour to influence the system's interpretations with unpredictable effects (see, e.g., Zaslow, 2002).

The second user-oriented criterion proposed is that of *scrutability* (Kay, 2000). The term scrutability is typically employed in user modelling to signify that every user's model can be inspected and altered by its owner. The goal is to enable users to determine themselves what is modelled about them and how adaptations based on their models will be conducted. In the context of this layer, the relevant dimension of scrutability that applies is the capacity of users to determine (i.e., inspect and control) how (or even whether) specific actions of theirs are interpreted by the system.

It is worth discussing at this point the notion of interaction between evaluation criteria. For example, making a system thoroughly scrutable may directly contribute to the system's predictability from the user's perspective. Although in the preceding example the interaction is contributory, we will later encounter cases where attempting to maximize one criterion will have potential adverse effects on others. It is very important to have a clear picture of such interactions between criteria when evaluating a particular adaptive system, and, if possible, to decide beforehand what types and levels of trade-offs between "competing" criteria are acceptable.

A summary of this layer is given in **Error! Reference source not found.**.

Table 3: Summary of Interpretation of the Collected Data Layer

Interpretation of the Collected Data (ID)	
Goal	Check that input data is interpreted correctly
Evaluation criteria	Validity of interpretations, predictability (of system's interpretations), scrutability (of system's interpretations)
Evaluation methods	Data Mining (see 4.1.3); Heuristic Evaluation (see 4.2.2); Play with Layer (see 4.3.1); Simulated Users (see 4.3.2); Cross Validation (see 4.3.3)

3.2.3. Modelling of the Current State of the "World"

This stage of the proposed model concerns the derivation of new knowledge about the user, the user's group, the interaction context, etc., as well as the subsequent introduction of that knowledge in the "dynamic" models of the IAS. There is a definite overlap between this stage and the interpretation of the input data; in fact, in several cases, there is no "second-level inference" in adaptive systems, which may simply go from interpreting the input data to representing those interpretations in an appropriate model. However, more often than not, IAS do employ second-level inference, mainly in the direction of relating the interpreted input to the current state of the dynamic models, as a basis for deciding the next "state" of those models. Inferences can be derived in many different ways ranging from simple rule based algorithms to Bayesian Networks, or Case-Based Reasoning systems.

The main evaluation criterion for this stage is *validity* of the interpretations/inferences. This refers to whether the inferences/interpretations reflect the actual state of the entity being modelled. Whereas, in many cases, this can be determined objectively and in a user-independent manner, this is not always true. For example, an IAS's inference on a user's interest in a particular piece or category of information (e.g., a tourist information system might infer user interest in visiting sites of historical interest), can only be judged on the subjective basis of the individual whom the inference concerns. In the context of recommender systems, validity is usually tested by n-fold cross validation (see Section 4.3.3) of a given dataset (Degemmis, Lops, & Semeraro, 2007; Berkovsky, Kuflik, & Ricci, 2008; de Campos, Fernández-Luna, Huete, & Rueda-Morales, 2009). While this setting makes it easy to compare and benchmark different modelling mechanisms and algorithms, such standardized datasets are not available in many other domains. In these cases, validity may be assessed through external criteria such as expert ratings (Suebnukarn & Haddawy, 2006) or prediction of a user's behaviour or performance (Yudelson, Medvedeva, & Crowley, 2008).

Beyond validity, it is argued that *predictability* and *scrutability* also need to be evaluated in this layer, although from a slightly different perspective than the one adopted in the previous layer. Specifically, predictability, in this case, refers to whether users are capable of predicting the system's modelling behaviour, given the system's interpretation of their actions. Similarly, scrutability, in this case, refers to the users' capacity to inspect and modify the user model itself (as opposed to the processes leading to its creation, or the ones involving its utilization). It is in fact this context of scrutiny and tailoring that the term is usually employed to convey.

While the above three criteria (and especially validity) are inarguably the most important ones for this stage, there are a number of lower-level criteria that address the modelling process in further detail. It is important to note that, in most cases, it only makes sense to

proceed with these criteria *after* reasonable levels of validity have been ascertained. The proposed lower-level criteria are: (a) comprehensiveness of the model; (b) conciseness of the model; (c) precision of the model; and, (d) sensitivity of the modelling process.

The first criterion, *comprehensiveness* of the model, is derived from information theory and is intended to identify the degree to which the IAS's model is capable of representing in its entirety the inferred/interpreted information about the entity being modelled. In other words, this criterion is concerned with how well the model can capture all the knowledge that is produced by the system within this particular adaptation stage (for instance, whether there are any properties that should be modelled, but are not modelled). Apparently, this is a criterion addressing the "structure" of the model and its representational power, as these relate to the inference process itself. Consider for example an adaptive learning system which offers learners, for every concept to be learned, a main description, a set of examples, and a set of self-tests to take. If the system's learner model is only capable of representing concepts as "known" or "not known"(or even a range of possible values between these two extremes) rather than modelling the learner's interactions in more detail, then it will be impossible to take advantage of the rich set of interactions available (and learning states possible) in further adapting the system to the user.

The second lower-level criterion, *conciseness* of the model, is "symmetric" to the first, and seeks to identify properties of the entity being modelled, which can be represented in the model, but cannot be inferred from interaction (and, thus, do not need to be modelled). This criterion is only relevant if the presence of such redundant "attributes" has adverse effects on the system's design or run-time operation (e.g., if the system's complexity is unnecessarily increased or run-time behaviour is impacted). Returning to the example of the adaptive learning system in the previous paragraph, consider the case where the user model is capable of representing learning progress in a fine grained way, including explicit representations of sub-elements of concepts, such as examples, tests, etc. If these sub-elements are not well structured in the system (e.g., not semantically distinguished within pages, necessitating that assumptions are made as to whether the learner has encountered them or not, in which order, etc.) then it will not be possible to populate the respective entries in the model dependably. This might be a problem if computational effort is expended in deriving the poorly substantiated entries, and, even more so, if the system's adaptation logic uses these values as if they were always present and dependable.

The criterion of *precision* of a model is again derived from information theory, and is concerned with the level of precision with which aspects of the user, context, etc., are modelled. For example, using a three-point scale to represent a person's knowledge of a given topic is certainly different than using an expanded seven-point scale, or a percentage. An alternative way to think about this criterion is that it is concerned with whether properties are modelled with enough detail. Whereas a high level of precision is, in general, a desired property of IAS, pursuing it may lead to redundancies, without necessarily increasing the comprehensiveness of a model. An important differentiating characteristic between the criteria of comprehensiveness and precision is that the first is mostly concerned with entire aspects of the entity being modelled, while the second addresses the "granularity" of the model and the level of precision that can be afforded.

Lastly, the criterion of *sensitivity* seeks to identify, on the one hand, how fast the modelling process converges to a comprehensive and accurate representation of the entity being modelled, and, on the other hand, the effects that fluctuations in the input data have on their respective models. Evidently, the desiderata are to: quickly arrive at a model that sufficiently represents the outside "world" (including the user), addressing in the process any "cold-start" problems (Schein, Popescul, Ungar, & Pennock, 2002); avoid "chase effects" that may result from the system's being too sensitive; and, avoid unnecessary latencies between an evident change in the modelled entity and the propagation of that change into the model. This

is a very delicate subject that needs to be approached with great care both in terms of adaptation design and in terms of evaluation. To better comprehend the complexity of the particular criterion, consider the example of a system that tries to "understand" whether the user's lack of interest in a previously appealing subject is temporary or the result of a more permanent shift in the user's interests.

A summary of this layer is given in **Error! Reference source not found.**.

Table 4: Summary of the Modelling the Current State of the "World" Layer

Modelling the Current State of the "World" (MW)	
Goal	Check that constructed models represent real world
Primary evaluation criteria	Validity of interpretations or inferences, predictability (of system's modelling behaviour), scrutability (of user model)
Secondary evaluation criteria	Comprehensiveness, conciseness, precision, sensitivity
Evaluation methods	Focus Group (see 4.1.1; 4.2.1); User-as-Wizard (see 4.1.2); Data Mining (see 4.1.3); Heuristic Evaluation (see 4.2.2); Play with Layer (see 4.3.1); Simulated Users (see 4.3.2); Cross-Validation (see 4.3.3)

3.2.4. Deciding upon Adaptation

During this adaptation stage the IAS decides upon the necessity of, as well as the required type of, adaptations, given a particular "state" (as the latter is expressed in the various models maintained by the system, or directly from input data).

Usually there are several possibilities of adaptation given the same user properties. Besides the way in which the system usually adapts, it is often possible to ignore the user model completely, or to use a single stereotype for all users. Furthermore, for most systems, there are even more adaptation behaviours possible. For instance, a product recommendation system might have inferred a strong preference for a specific product. It might now either recommend this product to the customer, only limit the possible selection to this product, indicate that there is a suggestion without naming it, or even recommend another product randomly. Comparing these alternatives might help to explore a kind of baseline that indicates what usual (non-intelligent) behaviour could achieve and whether adaptation really has advantages. As already discussed, one should be careful when using comparative analysis, especially if the "static" system compared against is a "without adaptation" version of the system being evaluated.

There is a very clear distinction between this stage and the next (see "Applying Adaptation Decisions" below). This separation can already be seen, for example, in the model of Oppermann (1995), where the inferential component (where, among other things, adaptation decisions are made) is separated from the efferential component (where adaptation decisions are applied). Again this is not necessarily a distinction that exists in practice; it is rather a way of facilitating the conceptualisation of the steps that are involved in the derivation and application of adaptation decisions, and which are often overlooked in evaluating, leading to questionable evaluation results. One "rule of thumb" that we propose for the separation between these stages is that decisions made at this stage are mainly at the semantic and

syntactic level of interaction; any further decisions made while effecting adaptation should belong to the lower syntactic, or to the lexical/physical level of interaction[6].

The goal in making this seemingly artificial distinction is to foster the separation of the adaptation theory (i.e., the foundation of the logic that drives adaptation) from decisions (made at design- or run- time) that represent a typical interaction design task, rather than a particular adaptation artefact. To return to our very first example again: A decision to guide learners would belong to this level; the same is true for more detailed versions of that decision, such as "to guide the learner by augmenting links in-place as they appear in the text". All other lower-level decisions (e.g., colour and adornments used to augment links) would belong to the next level though.

The primary aim of this evaluation step is to determine whether the adaptation decisions made are the optimal ones, given that the user's (and, more generally, the "world's") properties have been inferred correctly. We propose the following evaluation criteria for assessing the system towards this end: (a) necessity of adaptation; (b) appropriateness of adaptation; and (c) subjective acceptance of adaptation (i.e., does the user think that the adaptation is both necessary and appropriate?).

The *necessity* criterion is concerned with whether a decided upon adaptation is indeed required, given a specific interaction state, as this is represented in the system's various models. This criterion is often directly related to the theory underlying the system's adaptive behaviour, as it addresses the very point at which specific states of the system's "world model" are linked to (at least) high-level strategies for remedying identified problems, or capitalizing upon identified opportunities to support users in their interaction with the system (or with elements of their environment if the system's role is a mediating one). For example, in a system that seeks to automate commonly performed user tasks, the necessity criterion would need to be applied to all cases where the system identifies an action sequence the automation of which it believes will benefit the user.

Having established the need to adapt, one can then move on to the *appropriateness* of the decision made, i.e., is the adaptation decided upon one that can cater for the requirements posed by the current interaction context? For example, Nückles, Winter, Wittwer, Herbert, & Hübner (2006) observed expert behaviour and how it was influenced by the availability of additional knowledge to them. Essentially, they were able to demonstrate which adaptation decisions were taken by experts given different user models.

Finally, *subjective acceptance* of adaptation decisions refers to the user's perception of whether a decided upon adaptation is both required and appropriate. This criterion is complementary to the ones discussed thus far, in that it specifically urges evaluators to consider not only the objective dimensions of an adaptation decision, but also its direct effects as perceived by end users. It may well be the case that users are uncomfortable with a system decision, even if it is ultimately to their benefit (e.g., if it makes obvious a particular non-complimentary social trait of the user). Subjective acceptance is of particular importance when a lack of transparency may affect the user's trust in the system (Cramer et al., 2008) or compromise the user's privacy, e.g., in a group recommendation situation (Masthoff & Gatt, 2006).

When feasible and desirable, a number of more fine-grained user-oriented criteria can also be considered for this layer. To start with, *predictability of the system's adaptive behaviour*,

[6] The terms "lexical", "syntactic" and "semantic" refer to the three levels at which human-computer interaction occurs (Hoppe, Tauber, & Ziegler, 1986; Ziegler & Bullinger, 1991): The lexical level of interaction (also referred to as physical), which concerns the structure, presentation attributes, and actual behavior of the input / output interaction elements that make up the user interface; it is at this level that interaction physically takes place. The syntactic level of interaction, which concerns the structure and syntax of the dialogue between the user and the computer, through which the application semantics are made accessible to the user (e.g., specific interaction steps taken by the user, method of accomplishing tasks). The semantic level of interaction, which involves conveying the system functionality and domain-specific facilities to the end-user.

on the basis of its model of the "world" is an essential element in many domains of adaptivity. Complementary to predictability is the criterion of *scrutability of the system's adaptive behaviour*. As was the case in the preceding layer, these criteria reappear, but the perspective has again shifted to capture the portion of the adaptation process that is covered by this layer.

A criterion that is applicable in both this layer and the next is that of *breadth of experience*. Jameson (2008) argues that, especially in IAS that support the user in some form of information acquisition, the system's adaptive behaviour may prevent the user from experiencing the full range of items, products, functionalities, etc., that are available. This is related to *serendipity*, a criterion often applied for the evaluation of recommender systems (McNee, Riedle, & Konstan, 2006), and intended to convey the extent to which users make pleasant new discoveries when using the system. Jameson (2008) points out that a reduction of the breadth of experience is especially likely if the system relies more heavily than is necessary on an incomplete user model. Although this is quite possible, in the context of this framework this criterion is intended to identify decisions that have the described detrimental effects that are based on theoretically sufficiently populated and valid models.

It is noteworthy that we have here another clear example of interaction between proposed criteria. Specifically, some of the common methods used in systems to mitigate the diminishing of the breadth of experience, such as the systematic proposition of items that are not dictated by the current user model in a recommender system (see, e.g., Ziegler, McNee, Konstan, & Lausen, 2005), may have direct impact on the predictability of the system's behaviour. It is again recommended that evaluators explore such interactions, especially for novel criteria they add to the assessment of individual layers or the system as a whole, and that they ensure that the system's design priorities in this respect are reflected appropriately in the evaluation design.

A summary of this layer is given in **Error! Reference source not found.**.

Table 5: Summary of the Deciding upon Adaptation Layer

Deciding upon Adaptation (DA)	
Goal	Determine whether the adaptation decisions made are the optimal ones
Evaluation criteria	Necessity of adaptation, appropriateness of adaptation, subjective acceptance of adaptation, predictability (of system's adaptive behaviour), scrutability (of system's behaviour), breadth of experience
Evaluation methods	Focus Group (see 4.1.1; 4.2.1); User-as-Wizard (see 4.1.2); Heuristic Evaluation (see 4.2.2); Cognitive Walkthrough (4.2.3); Simulated Users (see 4.3.2); Play with Layer (see 4.3.1); User Test (see 4.3.1)

3.2.5. Applying Adaptation Decisions

This stage refers to the actual introduction of adaptations in the user-system interaction, on the basis of the related decisions. Although typically subsumed by adaptation decision making in the literature, this stage may be varied independently of the decision making process, e.g., to account for different adaptation strategies. More importantly, this stage usually "hides" a level of adaptation (i.e., the transformation of possibly high-level adaptation decisions to a "concrete" form experienced by the user), which only too often, and in several cases mistakenly in the authors' opinion, gets evaluated in tandem with the higher-level decision making stage.

The evaluation criteria that are applicable at this stage depend very much on the type of adaptation effected. In most cases, traditional evaluation criteria, such as *usability*, will be highly relevant (Gena & Weibelzahl, 2007). The identification of these criteria can only be performed on a case-by-case basis.

However, there are a number of adaptation-specific criteria that are largely independent of the type of adaptation and could be assessed at this stage. We propose that the following be considered as a minimum: *timeliness* of adaptation (i.e., is the decided upon adaptation applied in a timely manner - e.g., not too late?); *unobtrusiveness* of the adaptation (i.e., how obtrusive, or obstructive is the application of an adaptation, with respect to the users' main interaction tasks); and, *user control* over the adaptation (i.e., can the user disallow, retract, or even disregard an adaptation?). The last criterion is a specialization of *controllability,* which is discussed in detail in Section 3.2.7; it is repeated here explicitly to emphasize the role of the current layer in the users' perception of their control over the system. Furthermore, all of the aforementioned criteria can be thought of as directly contributing towards the *acceptance* of the adaptation by the user.

Criteria that have been suggested for prior layers and also have bearing on the application of adaptation decisions are *breadth of experience* and *predictability of the system's adaptive behaviour*. In terms of the former, assessment can address the extent to which the way in which adaptations are applied precludes (or makes less likely) that users will experience certain aspects of the system. In terms of the latter, assessment may address whether modifications (incurred by adaptivity) at the physical and syntactic levels of interaction are deemed predictable by the user.

The evaluation of this stage should be approached judiciously, and any related evaluation activity should be designed very carefully to measure only the relevant criteria. The difficulty in doing so arises from the fact that the users "experience" the grand total of the system's adaptive behaviour through the adaptations that are effected (and of which they are aware). Heuristic evaluations by experts in terms of usability criteria can help to detect issues with the

application of the adaptation decision at early stages of the development cycle (Carmagnola et al., 2008). An alternative approach that is more demanding on evaluation (and possibly also on development) resources, but enables the straightforward participation of end users in the evaluation activities, is the comparison of alternative manifestations of adaptation decisions. In such a scenario, two versions of the adaptive system would be in comparison, and they would differ only in how specific adaptation decisions are effected.

A summary of this layer is given in **Error! Reference source not found.**.

Table 6: Summary of the Applying Adaptation Decisions Layer

Applying Adaptation Decisions (AA)	
Goal	Determine whether the implementation of the adaptation decisions made is the optimal one
Evaluation criteria	Usability criteria, timeliness, unobtrusiveness, user control, acceptance by user, predictability (of system's adaptive behaviour), breadth of experience
Evaluation methods	Focus Group (see 4.1.1); User-as-Wizard (see 4.1.2); Heuristic Evaluation (see 4.2.2); Cognitive Walkthrough (4.2.3); User Test (see 4.3.1); Play with Layer (see 4.3.1)

3.2.6. Evaluating Adaptation as a Whole

The "piece-wise" evaluation of adaptation, as proposed in this paper, can provide valuable insight into the individual adaptation stages through which an IAS goes. However, what is still missing is the "big picture" – the evaluation of the primary adaptation theory (or theories). For example, the basis of adaptation in an adaptive learning system might be that guiding learners through the available material, decreases learning time and increases retention time of learned material.

To assert whether such high-level theories (or, seen from a different perspective, hypotheses) hold true, one needs metrics that transcend the layered evaluation of adaptation as this has been discussed so far. Such metrics must adequately capture the application- and adaptation- domains, to be able to more holistically assess the "success of adaptation". This role cannot be fulfilled by the stage-based evaluation criteria proposed in preceding sections, as these are "domain-agnostic", i.e., they make no assumptions, but also no provisions, for any particular application domain.

Browne, Norman and Riches (1990) have proposed that this problem be addressed by: (a) articulating and assessing against the system's objectives, and/or (b) assessing indirectly against the underlying theory. In the first case, the evaluation is centred around the identification of the objectives that the system aspires to attain (e.g., to speed up the user's interaction with the system, or to decrease the user's error rate, or to increase user satisfaction, etc.) According to Browne et al. (1990) many of the objectives of an adaptive system can be expressed as lists of purposes, which, in turn, can be loosely interpreted as the collection of "reasons" that led to the introduction of adaptation in the system, in the first place. Metrics and assessment methods can then be devised to measure the extent to which the stated objectives are met. These metrics might either be subjective, such as perception of and satisfaction with the system (Zimmermann & Lorenz, 2008) or may be objective, such as task completion time or number of steps required (Bontcheva & Wilks, 2005).

Following the above approach may not be equally straightforward when the success of the system in obtaining its objective is related only indirectly to the aspect of the user interaction that the system is attempting to improve. The means for attaining the objective may rest on an untested theory. For example, adaptation could be based on a theory that attempts to decrease error rates in order to increase user satisfaction. In order to test this theory it is essential that error rate data be collected (even though this does not reflect the objective of the whole system), and associated with evaluation results regarding subjective user satisfaction. In this case, error rates would serve as an indirect metric towards assessing the adaptive theory. Related to this concept is that of a "mediator variable" in statistics; the mediator variable is typically intended to concretize and/or operationalise the relationship between an independent and a dependent variable. When applying this approach evaluators are cautioned that it is sometimes very complicated to establish causal relationships between variables in an empirically rigorous manner (Green, Ha, & Bullock, 2010).

Apparently, the formulation/selection of metrics in both of the preceding cases is domain- and system- dependent. The establishment of such metrics needs to take place at design time, and their assessment must be planned well in advance, as there is a distinct possibility that related measurements may require monitoring aspects of the system's or the user's behaviour which are not part of the primary adaptation cycle. The selection of appropriate evaluation/assessment methods and instruments depends, naturally, on the very nature of derived metrics. For instance, one would approach in entirely different ways metrics related to interaction speed, from those related to user satisfaction, or retention of learning material. Section 5.1 discusses approaches that may be employed to identify and sufficiently specify domain-specific metrics for a system to be evaluated.

The discussion until now may have led readers to the conclusion that the assessment of interaction as a whole cannot be approached in a domain-independent way. However, this is not necessarily the case. If we accept that there are adaptation goals that are shared by IAS in different domains, then we could also formulate metrics that go beyond individual domains. For example, Weibelzahl (2003) proposes as a general goal of adaptation the simplification of the interaction process, and goes on to introduce the metric of *behavioural complexity* as a means of assessing against the stated goal.

A summary of this layer is given in **Error! Reference source not found.**.

Table 7: Summary of the Evaluating Adaptation as a Whole Layer

Evaluating Adaptation as a Whole	
Goal	Summative evaluation of the adaptation theory
Evaluation criteria	Specific for system's objectives or underlying theory
Evaluation methods	Heuristic Evaluation (see 4.2.2); Cognitive Walkthrough (see 4.2.3); User Test (see 4.3.1); Play with Layer (see 4.3.1)

3.2.7. Criteria Applicable in Most Layers

Beyond the criteria introduced for the individual layers there are some that apply to most, or even all layers. Namely these comprise *privacy*, *transparency* and *controllability*.

Privacy has been identified as a challenge to adaptive systems (Jameson, 2003) due to the potential tension between the use of personal data for personalization and the user's need for and concern of privacy (Kobsa, 2007). In fact, privacy is a complex issue that cannot be addressed by a single solution. All layers are affected by this issue. Starting with the *Collection of Input Data* layer, it may be necessary to evaluate whether users are willing to provide a certain type of information (Ackerman, Cranor & Reagle, 1999) or whether the data

is allowed to be collected under certain legislation (Wang, Chen & Kobsa, 2006). In regard to the MW layer it may be evaluated whether the information in the user model is stored in a secure way. In regard to the DA layer, it may be evaluated whether the adaptation may potentially disclose information about the user to other users.

On a similar note, the criterion *transparency* may need to be evaluated with respect to several or all of the layers. In general, it is desirable that a user of an adaptive system can understand why the system has made a particular adaptation or recommendation and how the system's adaptive mechanisms work (Jameson, 2003). Accordingly, depending on the system domain and application it may be important that the user is aware which information is collected (CID layer), which inferences are drawn (MW layer) or why a certain adaptation has been chosen (DA layer). Transparency is closely coupled though not identical to scrutability (Kay, 2000). A scrutable system allows users to inspect their user model and to change it, e.g., in order to remove inaccuracies. Mapped onto the layered approach it may be evaluated whether a user can undo or change system interpretations (ID layer), undo or change user modelling actions (MW layer), or undo and change adaptation decisions (DA layer). Depending on the IAS at hand, it may well be that scrutability subsumes transparency in those layers that it is applicable. Care should be exercised when making this assumption, however, since transparency is typically understood more expansively and is thus applicable to more layers than scrutability.

Controllability in this context refers to the user's perceived ability to regulate, control, and operate the product (Zhang, Rau & Salvendy, 2007). Users feel in control if the system behaviour can be strongly influenced by the actions of the user (Norman, 1994; Winter, Wagner & Deissenboeck, 2008). The term is sometimes also used to refer to the system property of the ability to move a system around in its entire configuration space (Ogata, 2009) which is obviously related but not identical to the user's subjective perception of being able to do so.

In the design of adaptive systems, the "ability of the user to determine the nature and timing of the adaptation" (Jameson & Schwarzkopf, 2002, p.193) is a key issue, because in many systems the adaptation is triggered implicitly by user actions and while some users want to control each aspect of adaptivity, others may have less desire to do so. Controllability can be evaluated in all layers, but is of particular importance for the following three layers: When *Modelling the Current State of the World* users need to feel in control that they can influence what the system thinks about them. When *Deciding upon Adaptation* users should be able to control which decision is taken. When *Applying Adaptation Decisions* users should be able to control how the adaptation is implemented. The specific system goals and domain may determine how important controllability is considered for system success. While in adaptive learning systems a high level of controllability seems to be desirable (Kay, 2001), it may be less so with recommender systems or agent-based systems (Trewin, 2000). As in the case of transparency, there are commonalities between the criteria of controllability and scrutability, and similar caveats apply as controllability does not necessarily imply a satisfactory level of scrutability, but the opposite is usually true.

Further to the above, there may also exist functional criteria that are of importance to the success of adaptivity and need to be subjected to user-based evaluation. For example, whereas algorithm complexity is something that can be studied independently (see, e.g., Domshlak & Joachims, 2007), it may be necessary to assess in real-world conditions the efficiency of a system as perceived by its users. Assessment of this kind of criteria may sometimes require the simulation of scale (e.g., in the number of users in the target population), which may be

addressed with hybrid approaches, such as the employment of simulated users (see Section **Error! Reference source not found.**).

4. Methods for the Formative Evaluation of IAS

Having presented the proposed evaluation framework for IAS, we now proceed to discuss a number of evaluation methods that can be used in conjunction with the framework. Returning to the differentiation between formative and summative evaluation, we would like to point out that, due to the nature of layered evaluation, this paper focuses almost exclusively on formative methods. This bias is due to the fact that, whereas summative evaluation methods are well established and in wide use, the same is not true of methods that are formative in nature and have been specifically tailored or developed to cater for the distinct nature of IAS. This section describes some of these methods and their application in the proposed framework.

Several publications already discuss and/or provide overviews of evaluation methods for adaptive systems (for example, Chin, 2001; Gena, 2005; Gena & Weibelzahl, 2007; van Velsen, van der Geest, Klaassen & Steehouder, 2008).

To start with, Chin (2001) presented a detailed discussion of factors that need to be considered when planning an empirical evaluation of an adaptive system (termed a "user test" in this paper). Chin (2001) placed emphasis on producing rigorous experiments that are well-controlled, use appropriate statistics, and are reported in sufficient detail. The discussion was restricted to summative evaluations (and is, in fact, an excellent guide for evaluators interested primarily in this type of studies). Qualitative methods were only briefly presented and their employment in adaptive systems not directly addressed.

Gena (2005) and Gena & Weibelzahl (2007) provide a comprehensive overview of evaluation methods for the adaptive web, derived from research in HCI. Inspired by Preece, Rogers, Sharp & Benyon (1994) and by Dix, Finlay, Abowd & Beale (1998), Gena (2005) classifies these methods into: (a) collection of users' opinions, (b) observing and monitoring usage, (c) predictive evaluation, (d) formative evaluation and (e) experiments and tests. This classification lacks clarity due to overlaps: for example, observing usage is often done while doing an experiment, and predictive evaluation is often of a formative nature. In later work (Gena & Weibelzahl, 2007) this is rectified and a distinction is made between data collection methods (including the collection of user opinions and user observation methods) and evaluation methods. Gena & Weibelzahl (2007) classified methods according to: (a) the factors the methods are most suitable to generate and evaluate (e.g., user satisfaction); (b) applicability conditions (e.g., a prototype and presence of expert evaluators); and, (c) their advantages and disadvantages. Both papers provide an overview of layered evaluation approaches, however the link between the general HCI methods discussed and the evaluation of layers in an adaptive system is not made.

Van Velsen et al. (2008) review user-centred evaluation studies of adaptive and adaptable systems. Rather than providing a framework for evaluation, they have taken a descriptive approach: mapping the current user centred evaluation practice, reflecting on its weaknesses and providing suggestions for improvement (e.g., the need to report think-aloud protocols in more detail than current practice). Most of the methods discussed would be qualified as data collection methods (as identified in Gena & Weibelzahl, 2007), rather than evaluation methods; of these, questionnaires were identified as the most popular method in user-centred studies of IAS.

In contrast to these existing overviews, this section provides an overview of methods specifically tailored or developed for the evaluation of IAS. This overview is not intended to be exhaustive; for more comprehensive accounts of general HCI evaluation methods readers are referred to Maguire (2001), and Gena & Weibelzahl (2007). As already discussed, the

focus will be on formative evaluation methods, and on relating these to the layered evaluation of IAS. Furthermore, methods will be related to phases in the life-cycle of system development, as this arguably largely determines their applicability/suitability, and, ultimately, selection.

In particular, this section distinguishes three development phases in which a layer may be evaluated: *specification*, *design*, and *implementation*. In contrast to the phases proposed by Gena and Weibelzahl (2007), the specification phase excludes the earlier period, in which the purpose and goals of a system may be unclear, necessitating investigations of the characteristics of users, tasks and contexts of use. The assumption made here is that such investigations, if necessary, have taken place prior to the specification phase. In contrast to the phases proposed by van Velsen et al. (2008), no distinction is made between high-fidelity prototypes and full systems. Arguably, formative evaluation is needed even when a system has been fully implemented. Perhaps even more importantly, most of the evaluation methods that can be used with a fully implemented system are also applicable with a high fidelity prototype.

The following sub-sections discuss and provide examples from the literature for the methods that can be applied in each of the above three phases for the different layers. As formative evaluation aims to inform and improve system design, this blurs the distinction between design and evaluation. Actually, in user-centred/participatory design, users are involved from the start, and design and evaluation go hand in hand. Therefore, methods suitable for the specification phase have also been included. **Error! Reference source not found.** provides an overview of the methods, categorizing them on the basis of when, how, and by whom the evaluation is done. It also shows which layers the method is particularly suitable for. This table will be explained in detail in the subsections below and in section 5.3.

To conserve space, the terms “input to a layer” and “output of a layer” will be used to refer, respectively, to the input and output data of the portion of an adaptive system’s functionality that corresponds to one of the framework’s layers (and may match a particular system component, although this is not necessary).

Table 8: Overview of formative evaluation methods for IAS, against a number of selection dimensions.

Method	When	How			By Whom	Which layers*
		Layer's Input	Layer's Output	Quality assessed		
Focus Group	Specification	Given	Produced	Opinions	Users / Experts	MW, DA, AA
User-as-Wizard				Criteria or Gold-standard	Users / Experts	MW, DA, AA
Data Mining				Gold-standard	Experts	CID, ID, MW
Focus Group	Design	Given	Given	Opinions	Users / Experts	MW, DA, AA
Cognitive Walkthrough				Criteria	Experts	DA+AA, Whole
Heuristic Evaluation				Criteria	Experts	Any
User test	Implementation (or Design and Wizard-of-Oz)	Given or Produced		Opinions or Criteria or Performance	Users / Experts	DA, AA, Whole
Play with layer		Produced		Opinions or Criteria	Users / Experts	Any
Simulated users		Produced		Criteria or Gold-standard	Simulated Users	CID, ID, MW, DA
Cross Validation	Implementation	Given		Gold-standard	Experts	CID, ID, MW

* CID = Collect Input Data, ID = Interpret Data, MW = Model the current state of the world, DA = Decide upon Adaptation, AA = Apply (or Instantiate) Adaptation.

4.1. Methods for the Specification Phase

Methods described in this section can be used when the general functionality that corresponds to a layer in the proposed framework has been decided, but no design exists yet. In other words, the system's input, and the desired kind of output for that stage are known, but the way in which the input will be transformed into output is not (or not fully).

Three methods are particularly useful in this phase: *focus groups*, the *user-as-wizard*, and *data mining*. Which of these methods is most suitable depends on the nature of the layer and the availability of data. If the task under evaluation is one that humans may be good at, then the user-as-wizard and focus groups methods are appropriate. If it is possible to obtain a dataset that maps input onto ideal output, then data mining may be appropriate.

4.1.1. Focus Groups

Focus groups are a type of interview conducted on groups. Participants provide their opinions on issues in an informal group setting, facilitated by a moderator (Krueger & Casey, 2009). This method is typically used early in the development process, to gather user requirements or obtain initial feedback on designs and prototypes. It produces rich qualitative data about what users want and (dis-)like. The informal setting encourages discussion. Normally, multiple

focus groups are held on the same issues to avoid bias due to participant selection and group dynamics in one particular group.

When focus groups are used in the specification phase, participants are told what kind of input the layer would have and may be given examples of this input. They discuss how the layer should produce its output. For instance, when considering the AA layer of an adaptive news website, participants may be given examples of the layer's input such as "emphasize football news, de-emphasize cricket news" (i.e., the adaptation decisions made by the preceding DA layer). They are asked to discuss how the (de-)emphasizing should be instantiated. A discussion may ensue of the relative merits of emphasizing through bigger fonts, re-ordering the news' list, adding star annotations, etc. **Error! Reference source not found.** provides additional examples of how focus groups can be used in the specification phase. The "Input" column shows what participants are told about the layer's input. The "Task (Question to group)" column shows what the participants are asked to discuss. Where available, examples from real studies are provided. If no reference is given, the example is hypothetical and included for illustrative purposes only.

Focus groups are suitable if humans are good at the task under evaluation. Often, this means that they are more appropriate for evaluations addressing the later layers in the adaptation process (i.e., MW, DA, AA). Even when participants seem capable of and are vocal in discussing how a layer should operate, results need to be used with caution. Participants' subjective opinions may well be wrong. This is expressed in the well-known design mantra "users are not designers and designers are not users".

For more hands-on information on how to run focus groups see (Krueger & Casey, 2009).

4.1.2. The-User-as-Wizard

The user-as-wizard is a method introduced specifically to provide a structured way for using humans to inspire the algorithms needed in an adaptive system. This method was first fully described by Masthoff (2006), though it had been implicitly (partly) applied before (e.g., Masthoff, 2004; Masthoff & Gatt, 2006; de Rosis, Mazzotta, Miceli & Poggi, 2006; Nückles et al., 2006). It integrates ideas from both contextual design and the wizard-of-Oz.

Contextual design is an ethnographic method, in which users are observed in their work place, to find out how they go about their work, in what environment, using which artefacts (Beyer & Holtzblatt, 1997). The idea is that users are the experts in their tasks and that observing them is better than asking them questions, as users' behaviour is often instinctive and their knowledge tacit. For example, Anderson, Boyle & Yost (1985) based their geometry tutor on observations of the strategies employed by teachers. However, observing experts in their normal setting is not ideal either, as the experts may use background and contextual knowledge that are not available to a system. Also, such studies are limited to situations that occur in the real-world setting. Finally, they are limited to the design of the system as a whole, rather than individual adaptation layers.

Wizard-of-Oz (see Section 4.3.1) is a technique used in user tests in which a system designer plays the role of the system. It has, for example, been used when developing dialogue systems that use speech recognition, to be able to evaluate the interactions without having to worry about the quality of the speech.

Table 9. Examples of using focus groups in the specification and design phases.

Phase	Layer	Input	Task (Question to group)
Specification	MW layer of an ITS, which infers the learner's emotional state from test results and sensor data.	The learner has answered 70% of questions correctly on the last two tests. He is leaning forwards.	What do you think the learner's emotional state is?
	AA layer of a news recommender, instantiating (de)emphasis.	Need to emphasize football news and deemphasize cricket.	How do you think the emphasizing/deemphasizing should be done?
	DA+AA[7] layers of a recommender, deciding what features to use to explain a movie's suitability (Tintarev & Masthoff, 2007).	None provided, participants used their knowledge about their own likes and dislikes and what was important to them.	How would they like to be recommended or dissuaded from watching particular movies
	DA+AA layers of a recommender, deciding and instantiating how recommendations are presented (van Barneveld & van Setten, 2003).	None provided. Participants were given some general background on what a TV recommender system does.	Users produced mock-ups of the way recommendations could be presented and explained.
	DA layer of an adaptive health information system, deciding what to tell the patient's close friends (Moncur et al., 2008).	Scenario: "Imagine that you are the close friend of someone whose baby was admitted to Neonatal Intensive Care after it was born recently".	Asked to do a 'card-sorting' exercise: given cards showing suggested information items, they were asked to reach a consensus on what heading to place each card under (e.g., "Essential Information", "Not needed").
Design	DA+AA layers: overall look of an adaptive public administration website (Gena & Ardissono, 2004).	None provided.	System mock-ups were provided. What do you (dis)like? How can it be improved?
	AA layer of a museum guide, selecting the character to present the narrative (Damiano et al., 2008).	Some background on what the role of character was going to be.	Four possible characters for the museum guide were shown. What are your opinions on the characters, and which is the best one?
	DA+AA layers of a recommender, deciding how to explain (Tintarev & Masthoff, 2008).	None provided.	Screenshots of different ways of explaining recommendations were provided. Which are best and how to further improve them?

[7] We use the notation XX+YY to refer to cases where two layers are evaluated in combination, a subject that we return to in detail in section 5.2.

The method consists of two stages. In the first stage, the exploration stage, participants take the role of the adaptive system, or, most frequently, of functionality that corresponds to a particular layer. This investigates how humans perform the task that needs to be performed. In the second stage, the consolidation stage, this understanding is consolidated by participants judging the performance of others.

Exploration Stage

Participants are given a scenario describing a fictional user and their intentions (task). Using fictional users is a well-known technique in user-centred design, where so-called personas are developed as examples of user classes (Cooper, 1999; Grudin & Pruitt, 2002). Similarly, scenarios—stories of a persona doing a task—are used extensively (Carroll, 2000). Personas and scenarios are also used in cognitive walkthroughs (Wharton, Rieman, Lewis & Polson, 1994; see below). Therefore, personas created for this type of study may be reused at other development stages.

In this stage, participants are given the task the adaptive system is supposed to perform. For instance, consider a scenario that describes 7-year-old Mary visiting a museum, and indicates that Mary likes horses, flowers, and the colour pink. A task given to participants could be to recommend three paintings for Mary to view. There is no need to instruct participants that they have to adapt. They will automatically base their recommendations on what they know about Mary. Crucial to the success of the method at this stage is finding out the participants' reasons for their decisions and actions, as this reflects: (a) what participants found important, providing criteria on which to judge adaptation; and, (b) how they went about the task, providing inspiration for the adaptation algorithm. The same observational methods as for a user test can be used (see Section 4.3.1). The above process may be repeated for several scenarios.

Consolidation Stage

The consolidation stage verifies the acceptability of the human performance and determines in what respects it can be improved. Participants should not have been involved in the exploration stage.

First, participants are given: (a) a scenario involving a fictional user and their intentions; and, (b) an associated task. The scenarios and tasks used are typically the same as in the exploration stage.

Next, participants are shown a performance on this task for this scenario. This can be a human performance (as from the exploration stage), or it can be a system performance (e.g., using an algorithm based on the exploration stage). For example, suppose a navigation support system needs to create a hierarchy of items of interest to the user. In the exploration stage, participants have produced such hierarchies. In the consolidation stage, participants are shown some of those hierarchies, and some hierarchies produced by an algorithm. Participants are not told whether the performance was by a human or system. They are then asked to judge the quality of task performance (in the example, how good the hierarchies are), potentially on multiple criteria. These criteria may be based on factors found to be important through observations of participants in the exploration phase (e.g., participants wanted hierarchies to be balanced in depth), or from input from system designers, or from indications in the literature (e.g., discussions on breadth versus depth in the literature). Again, as in the exploration stage, it is important to find out participants' reasons for their judgments. Similar observational methods can be used as in the exploration phase.

This procedure may be repeated for several task performances, presented in randomized order. Judgments of human performance may be interspersed with judgements of system

performance. Note that this resembles a Turing test (Turing, 1950), in that we could say that our system performs well, if participants judge it as well as they judge human performance.

Person & Graesser (2002), for instance, used a Turing test to evaluate the naturalness of dialogue moves of an ITS, finding that bystanders were unable to discriminate between dialogue moves of the ITS and a human tutor. However, depending on the layer's task, we may want it to outperform humans (e.g., when detecting patterns in a user's typing behaviour), or may still be satisfied with performance that is below human performance (e.g., when recommending books).

Examples of the use of the method and its stages are presented in **Error! Reference source not found.**. The "Input" column shows what participants are told about the layer's input, in the form of a scenario. The "Task" column shows what the participants are asked to do (i.e., producing the layer's output in the exploration stage, and judging the layer's output in the consolidation stage). The "Observational method" column indicates what method was used to find out why participants acted / made their decisions the way they did.

A limitation of the user-as-wizard method is that it is not suitable for tasks that humans are bad at. Basing adaptation algorithms on human performance is only sensible if humans perform well. Some tasks are inherently more difficult for humans than for computers. For example, humans tend to be bad at processing large amounts of data. For such tasks, they may have difficulty not just deciding what to do, but also judging performance. As in the case of focus groups, this means that this method is most suitable for the later layers of the framework (i.e., MW, DA, AA). Another limitation is that participants' judgments may not always correspond with what would be best for users. For instance, in a study of a medical reporting system, it was found that while doctors said they preferred graphs, they actually performed better with texts (Law et al., 2005). For this reason, a normal user test will still be needed. The user-as-wizard method is only intended as an initial step in the design process.

4.1.3. Data Mining

Data mining can be a very useful formative evaluation method in the specification phase if representative data is available showing which inputs should result in which outputs. Data mining techniques can inform the layer's design by discovering patterns; for example, which features of the input are important to predict the output accurately (Mobasher, 2007; Mobasher & Tuzhilin, 2009). There are three ways in which such ideal, gold-standard, output data can be obtained.

Firstly, the ideal outputs could be part of an existing dataset. For example, when designing a system component that predicts a user's movie rating based on their ratings for other movies, we can use the MovieLens dataset (Herlocker, Konstan, Terveen & Riedl, 2004). If this data includes a particular user's rating for the movie, then this will be the ideal output for the layer when it receives as input the other data for that user.

Secondly, the ideal outputs may be measured directly in a special study. For example, when designing an algorithm that predicts learners' knowledge from their behaviour, one could have learners interact with the system, gather behavioural data (the algorithm's input) and administer a test to measure the learners' knowledge (the algorithm's ideal output).

Table 10. Examples of using the user-as-wizard

Layer	Stage	Input (Scenario)	Task	Observational method
DA layer of a group recommender system (Masthoff, 2004), selecting a sequence of items adapted to a group of users. As part of this it has to aggregate individual ratings.	Expl.	John, Mary, and Adam are going to watch clips together. A table shows their liking for each of the clips.	Decide which five clips the group should watch.	Justify
	Cons.	Slight variation on the scenario above: we used "You and two friends (Friend 1 and Friend 2)".	Judge your and your friends' satisfaction if shown a particular sequence of clips. Repeated for three sequences.	Justify
DA+AA layers of a support system (Nückles et al., 2006).	Expl.	Experts interacted with a real layperson and were given information about their knowledge level.	Write instructional explanations of computer and internet issues in response to queries asked by the layperson.	Think-aloud.
DA+AA layers of a system convincing people to eat more healthily (de Rosis et al., 2006), deciding on arguments and message structure.	Expl.	A story about a fictional friend, with details about her personality, goals, habits, and healthy eating facts.	Construct a message to convince this friend to eat more healthily.	None
MW layer of a group recommender, modelling the effect on satisfaction of another member's emotion (Masthoff & Gatt, 2006).	Expl.	Think of somebody [who meets some relationship criterion]. Assume you are watching TV together. You are enjoying the program a bit.	Judge how it would impact your satisfaction to know that the other person is really hating/liking it.	None
MW layer of a persuasive system, modelling how a user's attitude changes when presented with an argument (Nguyen et al., 2007).	Expl.	Participants were told about Adam and his current position on nuclear power.	Judge how a particular argument would change Adam's position.	Justify
DA+AA layers of a navigation support system, deciding how to group items of interest to the user together in a hierarchy and how to name groups (Masthoff, unpubl.)	Expl.	Participants were given a set of items of interest to a user.	Construct a suitable textbook hierarchy to contain the items, inventing titles for chapters, sections etc.	Co-discovery
	Cons.	Hierarchies were shown, most produced in the Exploration Stage, others computer-generated.	Judge the hierarchy on given criteria. Explain what they disliked most.	Justify

Thirdly, a special study can be set up to *indirectly* measure the ideal outputs. For example, when designing a component that predicts learners' emotional state from sensor readings, one could have learners interact with the system, gather sensor data, and use human observers to annotate observed emotions over time. This differs from the user-as-wizard method, as the observers are not really performing the component's task: instead of deciding based solely on the input (sensor data), they may use information unavailable to the component, such as facial expressions. So, this can be used even when humans are not good at the task targeted by the evaluation, but are good at producing the desired output using richer input. When using this approach, at least two annotators are required and one should report to what extent they agree (e.g., using Cohen's kappa).

Error! Reference source not found. shows examples of the use of data mining in the specification phase. The "Gold standard" column shows the gold standard used (combinations of layer input and ideal layer output). The table also shows how this gold standard was obtained, as this is the most difficult aspect of using data mining. The availability of gold-standard output data is also used to evaluate designed and implemented systems, see cross-validation below (Section 4.3.3). For more detail on data mining for personalisation, see (Mobasher, 2007; Mobasher & Tuzhilin, 2009). For more hands-on information on how to use data mining see (Witten & Frank, 2005).

4.2. Methods for the Design Phase

Methods in this section may be applied when the design has been (partly) completed. Initially in this phase, ideas will exist of how different system components will work, which may have been illustrated through storyboards showing what output the components will produce given certain input. Later in this phase, full algorithms and/or graphical user interfaces (GUIs) will have been designed, providing clarity of how the system and its parts will work. It is assumed that no (full) implementation of the functionality corresponding to the layer exists yet. In addition to the methods described here, a user test (a method typically associated with the implementation phase) may also be applicable, by using a wizard-of-Oz technique (see Section 4.3.1).

4.2.1. Focus Groups

While focus groups would most frequently be used in the specification phase, this method is applicable in the design phase too. The main difference is that participants are shown the system's input and output for the layer under evaluation, and discuss the output's appropriateness. **Error! Reference source not found.** shows examples of focus groups in the design phase.

The main limitation of using focus groups in the design phase is that they gather subjective opinions only; as mentioned above in the user-as-wizard section, what people say they like might not be best for them.

4.2.2. Heuristic Evaluation

In a traditional heuristic evaluation, usability experts judge a system's user interface against a set of criteria. The most popular heuristics in usability testing are Nielsen's heuristics (Nielsen, 1994a): ten broad guidelines based on a factor analysis of common usability problems. When evaluating a layer of an adaptive system, the experts need to be given examples of the layer's input and resulting output. They also need appropriate heuristics.

Table 11. Examples of data mining and cross-validation.

Method	Layer	Gold standard	How the gold standard is obtained
Data Mining and Cross-validation	MW layer of a recommender that assigns users to lifestyles based on on-line behaviour (Lekakos & Giaglis, 2007). Learns classification rules.	Known lifestyles for a set of users plus these users' on-line behavioural data.	Had a portion of the population complete a psychographic questionnaire that allows them to be classified into lifestyle segments.
	MW layer of a speech recognition system that predicts the next user command based on past behaviour and context of use (Paek & Chickering, 2007). Learns decision tree.	Known speech commands given by users for whom we also have contextual and past data.	Collected data from existing users: speech commands (transcribed), times of commands, personal data contained on devices at the time of each command, what types of commands users had enabled.
	ID layer of an ITS which infers a learner's emotions at any given time from conversational features (D' Mello et al., 2008). Learns which dialogue features predict which affective state (regression).	Known emotional state for learners for whom we also have conversational data.	Collected conversational data in interactions with the ITS, and afterwards measured the learners' emotions (in 20s intervals) using videos showing screen content, facial expressions, and speech. This was done through self-rating, and rating by peers and trained judges.
Cross-validation	MW layer of a recommender that infers a utility function (able to decide on the best k items) from users' qualitative preference statements (Domshlak & Joachims, 2007).	Known ratings and therefore known preference orders for users for whom we have preference statements.	Uses EachMovie and MovieLens datasets which contain movie ratings and attributes/genres. As preference statements were unavailable, these were generated using a decision tree learning algorithm.
	MW layer of a search support system that infers a user's preference for topics from click patterns (Stamou & Ntoulas, 2009).	Known interest in topics for searchers for whom we also have click & relevance data.	Collected data from Google query stream. Users provided their general interest in topics for each query and rated the relevance of visited pages.
	MW layer of a spoken dialogue tutoring system, which infers learning from, amongst others, affective state (Forbes-Riley et al., 2008).	Known learning for learners for whom we also have perfect input data.	Measured learning through a post-test. For the training, corpora of learner interactions were used, which had been annotated with affective states, turn correctness and discourse structure.
	MW layer of a movie recommender system (Degemmis et al., 2007).	Known ratings for movies.	Uses EachMovie dataset which contains movie ratings and movie attributes.
	MW layer of a dialogue system, predicting a session's outcome (Horvitz & Paek, 2007).	Known outcomes of callers' sessions.	Used data from the legacy system.

Carmagnola et al. (2008) report a heuristic evaluation using the heuristics associated with Jameson's (2003, 2005) five usability challenges for adaptive systems: predictability and transparency, controllability, unobtrusiveness, privacy, and breadth of experience. These challenges have been proposed in this paper as criteria to be used when evaluating a specific layer, or the system as a whole (see section 3.2). Jameson linked these goals to frequently encountered properties of adaptive systems that may have detrimental effects in attaining these goals. Accordingly, he proposed compensatory and preventive measures. Later, Jameson (2009) extended the original five challenges into nine "usability side effects" of adaptive systems. This list no longer includes unobtrusiveness, but adds: need to switch applications or devices, need to teach the system, unsatisfactory aesthetics or timing, need for learning by the user, and imperfect system performance.

Magoulas, Chen & Papanikolaou (2003) proposed an integration of Nielsen's original usability heuristics with layered evaluation for adaptive learning environments. For each layer, they selected a subset of Nielsen's heuristics which they deemed particularly appropriate, and added more detailed criteria for these heuristics. For example, for the ID layer, they selected "Learner control" and "Error prevention". For "Learner control", they added criteria such as "the same content is presented in various formats according to the learning profile". The idea of a specific set of heuristics for a layer appears promising, and it makes sense to have more detailed criteria for heuristics in the context of a layer. However, the heuristics they selected for each layer and the criteria proposed do not always seem appropriate. In the example given, the criteria do not seem to address learner control directly, but rather automatic adaptation to the learner profile.

Error! Reference source not found. shows examples of how the criteria introduced in Section 3 can act as heuristics appropriate for layered evaluation. It also shows sample questions that can be asked about the layers to judge how well they perform on the criteria. The table is not intended to provide the definitive set of heuristics for the evaluation of IAS, though it provides a good starting point. It is a challenge for the IAS community to produce such a set, and, in particular, to base such a set on an analysis of common problems discovered in IAS and the layers of IAS (similarly to Nielsen's factor analysis of usability problems).

Heuristic evaluation can in principle be applied to every layer, as long as appropriate heuristics have been agreed upon. The experts need to have expertise in heuristic evaluation, need to understand the meaning of the particular heuristics and questions used, and need to understand the layer's input and output. However, experts are not real users, so results need to be treated with caution. In addition, trade-offs between different heuristics may be required (as already mentioned in the discussion of the criteria in Section 3). For example, making the system status more visible may reduce unobtrusiveness. So, depending on the task and domain, some heuristics may be of lower priority than others, and the relative importance of heuristics for a particular IAS may need to be considered.

For more information on how to conduct heuristic evaluations see (Nielsen, 1994a).

Table 12. Examples of proposed evaluation criteria and questions that can be used in heuristic evaluation

Criteria	Example Questions
Transparency / Comprehensibility	– Does the user know and understand what the system has captured (CID), interpreted (ID) and modelled (MW), and why; what adaptation decisions it has taken and why (DA); and how adaptation has happened (AA)?
Predictability	– Is the user able to predict what the effect of their actions will be on the system's beliefs (ID, MW) and decisions (DA)? – Is the adaptation not making the user experience too inconsistent? (AA) – Are users asked to approve major changes to the system's appearance/functioning? (DA, AA) – Does the system follow the conventions of applications the user normally uses? (AA) – Are adaptations done in a way that fits with user's expectations from the real world? (AA)
Privacy	– Is the user informed about the kind of data captured about them (CID), the type of inferences drawn (ID, MW), and the way this data is stored and used (ID, MW, DA)? – Is the user able to decide the kind of data captured about them (CID), inferences allowed to be drawn (ID, MW), adaptations shown (AA), which data is stored (ID, MW) and what it is used for (DA)? – Is personal data protected in a way similar to the real world? (ID, MW)
Controllability / Scrutability	– Can the user undo or change system interpretations (ID), user modelling actions (MW), adaptation decisions (DA)? – Can the user influence how adaptations are applied (AA), and how inferences (MW) and decisions (DA) are made, e.g., by setting parameters that control the system's behaviour?
Breadth of Experience / Serendipity	– Is the user still able to access material that the system thought was less suitable for them (AA)? – Does the system allow users to make unexpected pleasant discoveries, rather than restricting experience (DA / AA)?
Unobtrusiveness	– Are explanations of system's actions not disturbing the user unnecessarily and too often? (AA) – Is the user's approval of system actions not sought too often, when it is not really needed? (ID, MW, DA)
Timeliness	– Is the timing of system actions (e.g., messages) appropriately adapted to the users' activities and context? (DA, AA)
Aesthetics	– Are automatic changes to the system's appearance aesthetically pleasing? (AA)
Appropriateness / Necessity	– How necessary was the action the system decided upon? (this and the next question should be posed for individual actions, rather than collectively) (DA) – How appropriate was the action the system decided upon given the interaction state (and history) and the system's adaptive theory? (this question is not intended to assess whether the theory is valid, but whether the action is consistent with the theory's relevant premises) (DA)

4.2.3. Cognitive Walkthrough

A cognitive walkthrough (Wharton et al., 1994) focuses on learnability: usability experts work through typical user tasks, and decide for each action whether a novice user might encounter difficulties. They use the correct action sequence to accomplish each task. For each action, they keep four questions in mind: will users expect to do this, will they notice the control (e.g., button), will they recognize the control is appropriate for this step, and will progress be apparent once it has been used. This method is most suitable for the evaluation of layers that have direct or indirect effects on the GUI, i.e., the DA+AA layers or the system as a whole. For example, consider the evaluation of the DA+AA layers of an ITS which annotates lessons with traffic light icons based on whether the learner is ready to learn them (after a knowledge test). A cognitive walkthrough can be used to evaluate if a novice user will be able to select the optimal sequence of lessons to reach a particular learning goal.

Sometimes it may be possible to evaluate earlier layers of the framework, such as the MW layer. For example, consider the evaluation of the scrutability of a user model. If there is a GUI which allows users to modify their user model directly, then a cognitive walkthrough can be used to evaluate whether a novice user will be able to change a given user model to a particular desired state. If there is no such GUI, a cognitive walkthrough can be used only if: (a) there is a GUI to provide input to the user model (e.g., rate news stories); and, (b) the modelling algorithm has been designed such that a correct action sequence can be identified (difficult when Machine Learning is used).

Modifications to the method may often be required to suit the evaluation of an IAS. In particular, experts will likely need to be provided with multiple action sequences per task; after all, the system's behaviour may well change depending on the user. Also, a cognitive walkthrough (when applied unmodified) typically looks at the first time a user does a task, ignoring that the system may change over time, after learning more about the user.

Unfortunately, there is a complete lack of reported cognitive walkthroughs in the IAS community, and therefore, no table with further examples has been provided. For more information on how to run cognitive walkthroughs see (Wharton et al., 1994).

4.3. Methods for the Implementation Phase

Methods in this section can be used when a prototype of the system functionality to be evaluated has been implemented. This may be a limited prototype which can only deal with a subset of inputs, or a full implementation.

4.3.1. User Tests

Once the functionality corresponding to an evaluation layer has been implemented, it can be tested by real users. Typically, users are given well-defined tasks to do; hence *task-based* user test will be used to identify the most common type of user test. Measurements are made of users' performance (e.g., how fast they learn in an ITS) and opinions. Observational methods are used to identify the cause of problems. The main difficulty of testing an individual layer of adaptation is that it may be hard for participants to provide the kind of input required, necessitating the presence of special interactive facilities to support the process (alternatives include doing indirect user tests, or employing simulated users, as discussed later).

Other pitfalls for the empirical evaluation of adaptive systems have been noted (Chin, 2001; Masthoff, 2002; Gena & Weibelzahl, 2007; Tintarev & Masthoff, 2009), but these are not specific to the layered evaluation of adaptive systems and are therefore not repeated here.

Observational Methods

Different observational methods can be used in a user test, such as:

– *Thinking-aloud*. Participants are asked to verbalize their thinking while performing a task

(Ericsson & Simon, 1993; Lewis, 1982; Nielsen, 1993). Nückles et al. (2006) asked experts to think-aloud when deciding what explanation would be best, given the learner's knowledge level. D'Mello, Craig, Sullins & Graesser (2006) used a variant called emote-aloud: learners verbally expressed their emotions. Porayska-Pomsta et al. (2008) suggested asking learners to describe what they are thinking and feeling.

- *Co-discovery*. Participants work together with somebody they know well, and their naturally arising discussion exposes their thinking (O'Malley, Draper & Riley, 1984).
- *Retrospective testing*. Using an interview or questionnaire, participants report their thoughts after the task has finished, possibly while watching a video of their actions (Nielsen, 1994b). The latter is also called retrospective thinking-aloud, while thinking-aloud during the tasks is sometimes called concurrent thinking-aloud.
- *Coaching*. Participants are encouraged to ask questions when they encounter problems, help is provided and notes are made of these issues (Nielsen, 1994b).

Some changes to the observational methods may be needed when evaluating an IAS. For instance, when investigating usability it is normally stressed that participants are not to be aided (unless using coaching), and not to be asked direct questions during the task as these may guide them. However, when evaluating an adaptive system this may cause problems. For instance, users may not even notice the adaptation occurring, which may make it necessary to interrupt them, and ask them about it explicitly. For example, when evaluating scrutability, and participants fail to notice the scrutability tool (as happened in Czarkowski, 2006), it may be good to lead them to it (making a note to improve its visibility). Alternatively, adaptivity-related activities may be incorporated in the tasks to alleviate this problem.

Further to the above, the normal limitations of observational methods apply also when evaluating IAS. It is often claimed that both thinking-aloud and co-discovery may interfere with participants' cognitive processes, slowing them down and making them behave differently than they normally would, as also noted by Chin (2001). However, the impact may depend on how strictly Ericsson and Simon's (1993) principles for thinking-aloud are followed. Adhering to these principles, *classic* thinking-aloud aims at verbalisation without mental processing, only prompting by "keep talking", not establishing personal contact or directing the participant's attention. Usability studies often use a more relaxed approach, which may lead to mental processing and interference with task performance (Ericsson & Simon, 1993). Even classic thinking-aloud has in some studies been found to decrease task performance (van den Haak, de Jong & Schellens, 2003) and increase task duration (Hertzum, Hansen & Andersen, 2009), though it has also been found to have little effect on participants' behaviour and mental processes (Hertzum et al., 2009). In contrast, relaxed thinking-aloud clearly changed behaviour and increased perceived mental workload (Hertzum et al., 2009). A change of behaviour is even worse when evaluating an IAS, as it may influence the adaptation taking place. Based on the above, the classic variant of thinking-aloud would be preferable for evaluating an IAS. The coaching method clearly changes task performance as participants can ask for help, and may, therefore, be inappropriate for IAS.

Thinking-aloud also requires training, and is less natural than co-discovery and coaching. A study by van den Haak, de Jong & Schellens (2004) found that participants enjoyed co-discovery more than both concurrent and retrospective thinking-aloud. However, co-discovery may be less natural/suitable when a system is supposed to adapt to an individual user (unless a user model is provided, as in the indirect experiments discussed below). Thinking-aloud and retrospective testing may lead to participants justifying their errors, and being insincere. Retrospective testing may suffer from participants not being able to recall why they did things. However, van den Haak et al. (2003) found that concurrent and retrospective thinking-aloud protocols revealed comparable sets of usability problems. Given the reduction in task performance for concurrent thinking-aloud, they argued in favour of retrospective thinking-aloud, while noting that it may be less suitable for more complicated tasks. In contrast, van

den Haak et al. (2004) argued in favour of concurrent thinking-aloud, as it is less resource intensive than retrospective thinking-aloud (which requires twice the amount of time) and co-discovery (which requires twice the number of participants). They did not find a difference in task performance in that study.

The best observational method is likely to depend on the available resources (time and number of participants), the task type and complexity, the type of participants (importance of participant enjoyment and naturalness), and the importance of avoiding changes in participant behaviour.

Wizard-of-Oz Technique

If a layer's functionality has not been implemented yet, it may still be possible to do a user test by using a wizard-of-Oz technique (Gould, Conti & Hovanyecz, 1982). A human "wizard" (somebody from the design team) simulates the system's intelligence and interacts with the user through a real or mock computer interface. This technique is used for rapid prototyping when a system is too costly or difficult to build (Wilson & Rosenberg, 1988). The wizards tend to follow a precise script. Participants are typically unaware that a wizard is used, and believe the system is fully implemented.

Wizard-of-Oz has been used in the evaluation of adaptive systems for a long time. For example, Maulsby, Greenberg & Mander (1993) used wizard-of-Oz to prototype an intelligent agent. Several recent UMUAI papers report on wizard-of-Oz studies (Miettinen & Oulasvirta, 2007; Batliner, Steidl, Hacker & Nöth , 2008; Damiano, Gena, Lombardo, Nunnari & Pizzo, 2008; Conati & Maclaren, 2009). For example, Miettinen and Oulasvirta (2007) used wizard-of-Oz to simulate the system functionality that corresponds to the CID / ID layers: sensors were simulated by human codings of data. In a layered evaluation, wizard-of-Oz can also be useful to simulate layers preceding the one being evaluated, to ensure these work perfectly and to enable the evaluation of a layer in isolation. A wizard could also help users to provide input for a layer that has no user interface normally.

As noted by Walker, Rummel & Koedinger (2009), wizard-of-Oz is impractical for large-scale research as it creates uncertainty as to whether different facilitators acting as wizards may have different effects.

Play-with-Layer

Play-with-layer is a variant of a user test in which participants are not given tasks, but allowed to freely explore the system or layer. They freely input data as if coming from the preceding layer in the adaptation process, and judge the output. There are several ways of judging a system's behaviour for a particular layer. Firstly, it can be judged against criteria. Secondly, a questionnaire or interview can be used to obtain participants' opinions. Finally, it may be possible to use objective measures, for example the frequency of occurrences of certain events, such as adaptations.

Indirect User Test

A problem with using a user test for an *adaptive* system is that adaptation takes time, often too much time to be able for the system to adapt during a typical one-hour experiment. One solution is to focus on evaluating the later layers in the framework, with the user model provided (by, or to, the participants). When the user model is provided to the participants, this comes down to an *indirect* user test. In contrast, standard user tests will be called *direct*.

In an indirect user test, participants perform the task on behalf of somebody else, rather than for themselves. This allows the evaluator to control the characteristics of the person for whom participants perform the task, avoiding the time delay otherwise needed for initializing and populating the user model from actual user interactions with the system. Importantly, an indirect experiment also ensures that the input to a layer is perfect, making it very suitable for

layered evaluations. George, Zukerman & Niemann (2007) used an indirect experiment because they wanted to focus on a particular behaviour of the system that did not always occur and wanted to remove extraneous factors from the evaluation. Indirect user tests are less natural for participants, and the results may therefore be less reliable.

Error! Reference source not found. shows examples of both standard user tests (task-based, direct), and indirect and play-with-layer variants. It shows the input of the layer, the task performed by participants (for standard and indirect user tests), the measurement and observational methods used, and the criteria that the layer is evaluated against.

For more information on how to conduct, design and report user tests see (Robson, 1994; Dumas & Loring, 2008).

4.3.2. Simulated Users

A general problem with user tests is that they tend to be costly in both financial and temporal terms. This may be further hampered by difficulties in recruiting a sufficient number of users. The situation is even worse when evaluating an adaptive system. Adaptation takes time, so a user study may need a long duration or even be longitudinal, with users taking part in multiple sessions. It may be hard to get users that can participate long enough for the adaptation to be fully tested. Recruitment is further complicated by the need for many different types of users to fully measure the impact of adaptation, the accuracy of user modelling, etc. In addition, comparatively more formative testing is probably needed in IAS, as they tend to contain intelligent algorithms, and many adaptation alternatives to compare. Finally, when evaluating an individual layer, it may be hard for users to provide the layer's input. For example, when evaluating the DA layer, it may be difficult for users to provide a correct user model. Or, as we have seen, even when it is possible for the user to provide the input, special interaction facilities may need to be implemented for this purpose. For these reasons, the simulated users method is based on computational models of users instead of real users.

In usability testing, model-based testing has been proposed as a way to quickly test systems without the need for real users. Methods such as GOMS (Card, Thomas & Newell, 1983) are used as a basis for implementing simulated users (e.g., using a probabilistic model).

Murray (1993) proposed the use of simulated students for formative ITS evaluation. This method has since been used by many ITS researchers (e.g., VanLehn, Niu, Siler & Gertner, 1998; MacLaren & Koedinger, 2002; Millán & Pérez de la Cruz, 2002; Guzmán, Conejo & Pérez-de-la-Cruz, 2007). It has also been used in the evaluation of other types of adaptive systems. **Error! Reference source not found.** shows example studies. It shows the layer's input produced by simulated users, the measurements taken, and the criteria on which the layer is evaluated.

The advantage of using simulated users is that different aspects of adaptation can be tested rapidly, and that the system inputs for the different layers can be controlled. The main problem is that the models used for building the simulated users are likely to be based on the same assumptions that underlie the adaptive system's design. What if those assumptions are wrong? For example, the simulated voices used for evaluation by Chickering & Paek (2007) are a subset of those used to train the baseline model. So, what if these simulated voices were unrealistic? A second issue is that modelling static user behaviour differs from modelling adaptive user behaviour. A model that accurately captures user behaviour when the system is static, does not necessarily accurately predict how users will behave when a system adapts. Finally, despite their usefulness in formative evaluations, simulated users will not be able to provide qualitative feedback, or provide subjective opinions on vital aspects of the system (e.g., aesthetics, feeling of trust). We therefore advocate using simulated users initially to gain rapid insight, and reverting to real users to validate findings. Indeed, most papers mentioned in **Error! Reference source not found.** report on additional studies with real users to either

validate the simulation models or to validate the findings of the simulations (Masthoff, 2002; Masthoff & Gatt, 2006; Guzmán et al., 2007; Hollink, van Someren & Wielinga, 2007).

Table 13. Examples of standard user tests (task-based and direct), and indirect and play-with-layer variants.

	Layer	Input	Task	Measurement	Observational Method	Criteria
Standard (task-based, direct)	MW layer of a recommender system	Produced by participants	Convince it you hate cricket.	Do they succeed and if so, how quickly? What causes problems?	Co-discovery	Transparency Scrutability
	DA+AA layers of a recommender deciding how to explain (Tintarev & Masthoff, 2007)	Participants set their own user model, via a specially made GUI.	Decide how much you like a movie.	Ratings of the explanations on various criteria	Justification	Effectiveness Persuasiveness
Indirect	DA+AA layers of an ITS that annotates lesson links	Told about a learner, and that the system had adapted.	Select a lesson to suit this learner.	Do they enjoy using the system, trust it, make appropriate and fast decisions? What causes confusion?	Co-discovery	Satisfaction, Effectiveness, Efficiency Trust
	DA+AA layers of a museum guide, which decides what to tell the user (Goren-Bar et al., 2006)	Told about a visitor, and shown videos of aspects of interaction with two guides.	Rate guides on aspects, pick best.	Justifications were analysed for statements related to the criteria. Analysed relation between personality and preferences.	Justification	Acceptability Transparency Usability Memorisability
Play with Layer	CID layer of a news recommender, deciding what the user looks at	Participants look at different parts of the screen.		How accurately and fast it picks up what they look at. Requires a GUI showing the layer's output.	None	Accuracy Efficiency
	ID layer of a recommender deciding interest based on what users look at	Participants position the mouse on items they look at.		Do participants agree with the inferred interests?	Retrospective	Accuracy Acceptability
	DA layer of recommender, deciding music based on users present and moods (Masthoff et al., 2007)	Participants set users' music preferences and simulate users entering and exiting.		Simulator shows the individuals' mood based on music played so far. Participants judge the decisions.	Justification	Effectiveness
	Overall experience of a museum guide (Stock et al., 2007)	Participants use the guide in a real museum setting.		Questionnaires	Retrospective	Ease of use Intention to use Involvement

Table 14. Examples of the use of simulated users.

Layer	Input (through Simulation)	Measurement	Criteria
DA layer of an ITS, deciding what word-pair to teach next (Masthoff, 2002).	Answers to practice items produced by simulated students, based on models of learning proposed in the literature. Simulations with varying models and parameter values.	How many correct responses the simulated learners get on average on a test, for different variants of the DA layer.	Effectiveness
DA layer of a Group Recommender, deciding which music item to play next (Masthoff & Gatt, 2006).	Affective state produced by simulated users. Simulations with varying parameter values.	How the simulated users' feel at any moment based on decisions made by different variants of the DA layer.	Effectiveness (to keep individuals satisfied)
MW layer of an ITS, inferring knowledge based on replies to questions (Guzmán et al., 2007).	Answers to test items produced by simulated students with known prior knowledge levels.	Comparing real knowledge with inferred knowledge. Measuring time.	Accuracy Efficiency
MW layer of a navigation support system, which divides a website's pages on the basis of user logs into sets that correspond to navigation stages (Hollink et al., 2007).	Navigation log files produced by simulated navigators (finite state automata modelling transition between navigation stages).	How often did the algorithm discover the right number of stages?	Accuracy
DA layer of a cognitive assistance system, deciding when to assist (Serna et al., 2007).	Actions and mistakes when performing a cooking task by simulated people with Alzheimer's disease. The simulation model is parameterized according to the different stages of the disease.	Can measure impact of assistance provided on number of mistakes made. Not really covered yet in this study.	Effectiveness
MW layer in a dialogue management system, personalising a baseline model to a voice (Chickering & Paek, 2007).	Speech commands produced by simulated voices with varied values for parameters.	Compared accuracy of different strategies for personalizing the model.	Accuracy

4.3.3. Cross-Validation

This method is appropriate for validating the accuracy of a layer's output if there exists a gold-standard: representative data showing which inputs should result in which outputs (see Section **Error! Reference source not found.**). The data is split into two parts: one part (called the training data) is used to inform the design of the system's functionality for a given layer. The other part (called the test data) is used to verify the accuracy of the (potentially implemented) design. To avoid accidental effects caused by the way the data is split, more rigorous forms of this approach tend to be applied, such as k-fold cross-validation (Kohavi, 1995): the data is split into k segments, and at any time k-1 segments form the training data, with the remaining segment acting as test data. This is repeated k times, with each segment in turn acting as test data.

Cross-validation is by far the most frequently used method in UMUAI papers of the last three years: it was used 20 times in the period 2007-2009 (compared to only 12 times in all the preceding years). This effect, however, may be partly due to the special issues on Data Mining and Personalization (Mobasher & Tuzhilin, 2009) and Statistical and Probabilistic Methods for User Modelling (Albrecht & Zuckerman, 2007). **Error! Reference source not found.** shows examples of cross-validation. Note that components constructed on the basis of results derived from data mining are normally evaluated using cross-validation.

There is a question about whether this method has a place in this paper, given our stated emphasis on formative evaluation. The method in itself is perfectly valid; however, evaluators that use it tend to only apply this method and then report on the accuracy of the evaluated component. Therefore, evaluations based on this method tend to be completely summative, without any formative insights. In our opinion, this does not have to be the case. The accuracies achieved tend to be quite far from 100%, and one wonders whether it would not be possible to analyse in what kind of cases the aspect evaluated is sub-optimal, so that at least some insight is gained into when it works well and when it needs improving. Another limitation of the method is that it only investigates accuracy (be it in all its forms, such as MAE, precision, recall, ROC) and sometimes efficiency, and there are many other criteria that may need evaluating. Finally, this method's need for gold-standard output normally makes it unsuitable for the DA and AA layers.

For more information on how to use cross-validation see (Witten & Frank, 2005).

5. Using the Framework

Having discussed the framework itself and formative evaluation methods that can be used in association with it, in this section we turn our attention to practical issues related to the employment of the framework. Firstly, we discuss how the application domain and type of adaptation employed may affect evaluation, and specifically the selection and operationalisation of assessment criteria. We then concentrate on the evaluation of layers in combination for the needs of particular systems and evaluation studies. This is, finally, followed by a synthetic view over the evaluation methods presented above, offering preliminary guidance for selecting a method (or methods) for specific evaluation settings.

5.1. Considering the Application and Adaptation Domain

Section 3.2 discussed a number of criteria that can be used when employing layered evaluation. These were selected on the basis of their generality and wide applicability, and are, in their majority, layer-specific. However, for all but the most trivial cases, there will be attributes of adaptation that are "cross-cutting concerns" over more than one (or even all) layers (see, for instance, the criteria proposed in section 3.2.7, or the criterion "breadth of experience" argued to be applicable both when deciding upon, and when applying adaptations – sections 3.2.4 and 3.2.5 respectively). Often, what these attributes are depends on the

application domain and the type(s) of adaptation supported. Their identification and operationalisation is not always a straightforward task, but the literature provides some guidelines that can assist towards this end.

One approach which can be used to guide the selection of criteria comes from Browne et al. (1990) who propose that a number of "metrics" be defined to assist in the design and evaluation of adaptive systems. Totterdell & Boyle (1990) provide a more detailed account of how these metrics can be used to drive the assessment of adaptation. Note that the word "metrics", as used in the preceding publications does not necessarily refer to measurable indices in a system, but rather operationalised discrete elements of the system's adaptive behaviour. Of the proposed metrics, some are of direct relevance to the discussion here (Browne et al., 1990):

- *Objective Metric*: captures the objective of the adaptive system (e.g., decrease error rate).
- *Theory Assessment Metric*: required when the success of the system in obtaining its objective is related only indirectly to the aspect of interaction that the system is attempting to improve (e.g., increase user satisfaction through reduced error rates).
- *Trigger Metric*: describes the aspect of user interaction on which the adaptation is based.
- *Recommendation Metric*: provides a description of the output of the theory-based part of the system (i.e., the adaptation decisions made by the system).

Totterdell & Boyle (1990) argue that by specifying and assessing these metrics in relation to one another, one can answer many questions about the functioning of an adaptive system. It is further argued here that the Objective- and Theory Assessment- metrics in particular, can serve as a guide for defining criteria that permeate the evaluation of individual layers or the system as a whole. Consider, for instance, a system that controls temperature and lighting in a house. For such a system, the Objective Metric may be associated with the automatically achieved comfort level of the inhabitants. The Theory Assessment Metric would then possibly address the effort levels that the inhabitants have to exert to attain the desired temperature and lighting settings, in relation to the system's initiative in modifying these settings. Note that these are but the first steps towards an evaluation design; these high-level metrics would then have to be broken down to measurable quantities that, in turn, can be derived through the application of selected evaluation methods and data collection instruments.

A second approach which can be used to guide the selection of criteria, and is along similar lines to the specification of metrics, is to focus on the dimensions of adaptation in a system, including its determinants and constituents, to arrive at the operationalisation of attributes that need to be assessed during evaluation. Knutov et al. (2009) identify six questions that, when answered, can provide a reasonably complete definition of adaptation in a system, as well as the ways in which they relate to each other (**Error! Reference source not found.**)[8]:

- What can we adapt? (What?)
- What can we adapt to? (To What?)
- Why do we need adaptation? (Why?)
- Where can we apply adaptation? (Where?)
- When can we apply adaptation? (When?)
- How do we adapt? (How?)

There are apparent correspondences between the metrics proposed by Totterdell & Boyle (1990) and the questions/dimensions put forward by Knutov et al. (2009). Perhaps the most important such correspondence is that between the Theory Assessment metric and the Adaptation goals (Why?), which is usually what an evaluation of an IAS sets out to assess in

[8] Knutov et al. (2009) restrict their analysis to Adaptive Hypermedia Systems, but the questions and their interrelations are arguably more generally applicable to most classes of IAS.

the first place. It is recommended that evaluation activities start from this very dimension to define measurable criteria for individual layers and the system as a whole. The integration of such criteria into an evaluation process driven by the proposed framework can take place from two complementary perspectives: (a) evaluators can specify the layers for which the defined criteria are relevant and incorporate them into their evaluation design; (b) the criteria, when they represent cross-cutting concerns, may also determine what combinations of layers (a subject addressed in the subsequent subsection) may be addressed to get a more holistic picture of the system.

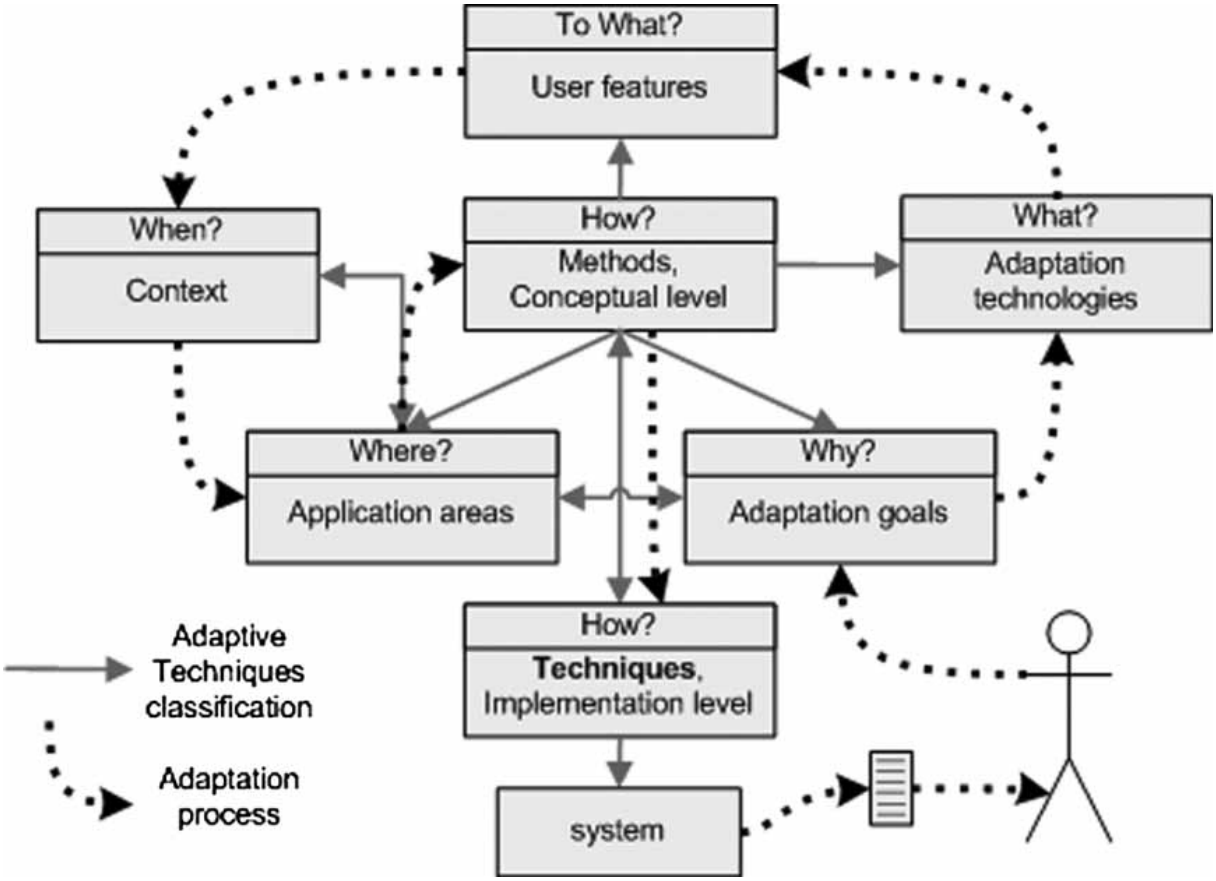

Figure 8: Classification of Adaptive Hypermedia methods and techniques, adaptation process highlights (Knutov et al., 2009)

Both propositions put forward here are intended to facilitate the process of formalizing the underlying design decisions in an IAS, so as to enable the derivation of the domain-specific criteria that will be used to assess these decisions. Of utility in this context may be other evaluation frameworks that propose complementary or alternative approaches to layered evaluation, and are discussed in Section 6.3.

5.2. Evaluating Layers in Combination

When presenting the evaluation layers, it was often remarked that evaluating them in isolation may not be feasible due to the nature of adaptivity in the system, the system's architecture, etc. In addition to such practical considerations, one may also have to observe organizational and resource constraints that may apply in the evaluation. For instance, a system may be sufficiently complex that evaluating each layer in isolation would require an amount of resources not readily available. When such constraints exist, or when assessment criteria need to be evaluated across layers as discussed above, it may be necessary to evaluate layers in

combination. This section discusses potential combinations of layers and considerations for their employment.

Starting from the end of the adaptation process, a combination that is often made in the literature is between the layers of deciding upon (the type of) adaptations, and the layer of effecting the said adaptations in the interactive front end. This combination is often motivated by the fact that most adaptive systems do not support alternative manifestations of adaptation decisions at the syntactic and lexical levels of interaction. For instance, an adaptive learning system usually supports only one way of denoting links are "ready to read". Although, in general, this combination is a reasonable one, evaluators should be careful when drawing conclusions about a system's adaptive theory from results thusly derived. This is especially true in the case that results are negative, since this could be attributed either to a faulty hypothesis serving as the basis of adaptation, or to an inappropriate incarnation of the adaptation decision at the physical level of interaction. This may be the case, for instance, with the results reported by Brusilovsky et al. (2001), where the authors ensured the validity of the user model, and concluded that the identified problems must lie with the adaptation theory – but did not separately check whether alternative manifestation of adaptive navigation support might have led to better results. At the opposite end of the spectrum, even if the IAS does distinguish between the two layers, it is possible to treat them jointly in terms of evaluation by: (a) enumerating all the possible concrete manifestations an adaptation decision may have, and (b) treating each decision - concrete instance pair as a distinct decision.

Another combination often made in the literature merges together the first three layers of the proposed framework, treating the collection of input data, its interpretation, and the modelling of the resulting knowledge as a single step or a single stage in the adaptation process. Again, this is in many cases a reasonable combination, but may suppress the true origin of identified problems. Consider the case of a personalized museum guide, in which visitations of artefacts in the museum's physical space are used to infer the visitors' interests in different styles, epochs, artists, etc. If the evaluation of the first three layers in combination shows that the user model only poorly represents the users' real interests, what should that be attributed to? The system's component that determines a person's position and direction of sight in the museum's rooms? The algorithm that translates a series of positions into "visits"? The assumption that visitors will only stand in front of artefacts that fall within their interests? The extrapolation of common characteristics between the visited artefacts? In an evaluation that merges together the first three layers, such questions may be impossible to answer with any certainty.

A combination that is potentially less challenging than the aforementioned one merges together only the first two layers of the proposed framework, namely the collection of input data, and its interpretation. This can be entirely straightforward in situations where the interaction data assembled is unambiguous, and/or represents the entirety of data observable by the system. In such cases, the only processing that occurs and may, therefore, result in errors, is concentrated in the interpretation of the collected data. If, however, this premise does not hold, this combination is susceptible to the same kind of problems discussed above.

It should be noted that by adopting two of the layer combinations discussed above, namely treating the first three of the proposed framework's layers as one and the last two likewise, we effectively arrive at the two-layer decomposition proposed by Karagiannidis & Sampson (2000). Employment of the two-layered evaluation approach is a major step forward from traditional practices that make no attempt at assessing individual adaptation steps, and could be considered the most minimalistic decomposition plausible for evaluating adaptation.

In summary, combining layers is a reasonable approach under certain circumstances, and possibly the only feasible one in some cases. When employing it, however, researchers and practitioners should exercise additional caution when: (a) using criteria that are meant for the evaluation of individual layers (and whose semantics may be diffused when merging layers);

and, more generally, (b) planning the evaluation to prevent the occurrence of unattributable effects. All potential difficulties that arise when merging layers can be traced back to the fact that the individual layers still exist, but are "hidden" (as are their effects on adaptation) from the perspective of the evaluator. A thorough understanding of this fact and its repercussions is, in the authors' opinion, a prerequisite for the successful application of the layered evaluation approach with combined layers.

5.3. Selecting Evaluation Methods for Layered Evaluation

In the planning of evaluation studies, once decisions have been made regarding the layers (or their combinations) that need to be assessed, and the criteria this will be done against, the next issue to tackle is the selection of the evaluation methods most appropriate for the evaluation settings. The presentation of methods for the formative evaluation of IAS in Section 4 has adopted the explicit assumption that the most appropriate evaluation method(s) in a given situation will depend on the development phase. Other factors to be considered include who will be involved and which data is available.

From the overview of methods it is clear that the evaluation can either involve users, experts, or simulated users. Users are the most realistic participants, as they are the ones who will end up using the system. Experts may be required when the layer's input and/or output is difficult to understand for ordinary users (e.g., for an IAS using a decision-theoretic model to decide upon adaptations). Experts may also have a better understanding of evaluation criteria (as required for example for heuristic evaluations and cognitive walkthroughs). Simulated users may allow for rapid and controlled testing of multiple alternatives.

Evaluation methods also differ in terms of the input and output data for the component(s) evaluated in each layer:

- *The layer's input*. The input data to the component(s) that embody the functionalities that a layer is intended to assess can be either given to the participating end users or experts, or decided by themselves. This could be input that is normally gathered over a long period of time, for example a user model that has been built up over a period of weeks. Allowing the participants to decide the input may require the development of special interaction facilities for this purpose, as most layers will lack this.
- *The layer's output*. Similarly to the input, we can either provide the output data of the component(s) corresponding to a layer to the participants, or let them produce that output themselves. Presenting such output may require effort, as most components involved in the adaptation process will not normally have a front end (interactive or otherwise). Another problem is that it may be hard to differentiate between outputs intended to be assessed at two layers (e.g., it may be hard to consider the outputs of the Apply Adaptation and Decide upon Adaptation layers separately).
- *A method for assessing the output's quality*. We can use subjective opinions, judge the output on criteria, or compare the output with a gold standard (the ideal output for the corresponding input).

Error! Reference source not found. (see Section 4, page **Error! Bookmark not defined.**) provided an overview of the methods discussed in this paper and how they differ on these aspects. **Error! Reference source not found.** puts together a set of rules of thumb that evaluators can follow, summarizing the discussion and propositions made in Section 4. The diamonds (◊) indicate questions that guide the selection of methods. For example, the first question is what the development stage is, and depending on the answer different methods apply. If it is the specification stage, then data mining, user as wizard, and focus group are possibilities. Which of these is best depends on further questions. For example, the later two methods are only suitable if it is a task humans are good at. Note that the organization in **Error! Reference source not found.** is only partial, and is intended to facilitate decision making, but not necessarily to be the sole driving force in this respect. After all, the best

method will also depend on the criteria one wants to evaluate. For example, the simulated users method is not suitable to evaluate subjective acceptability.

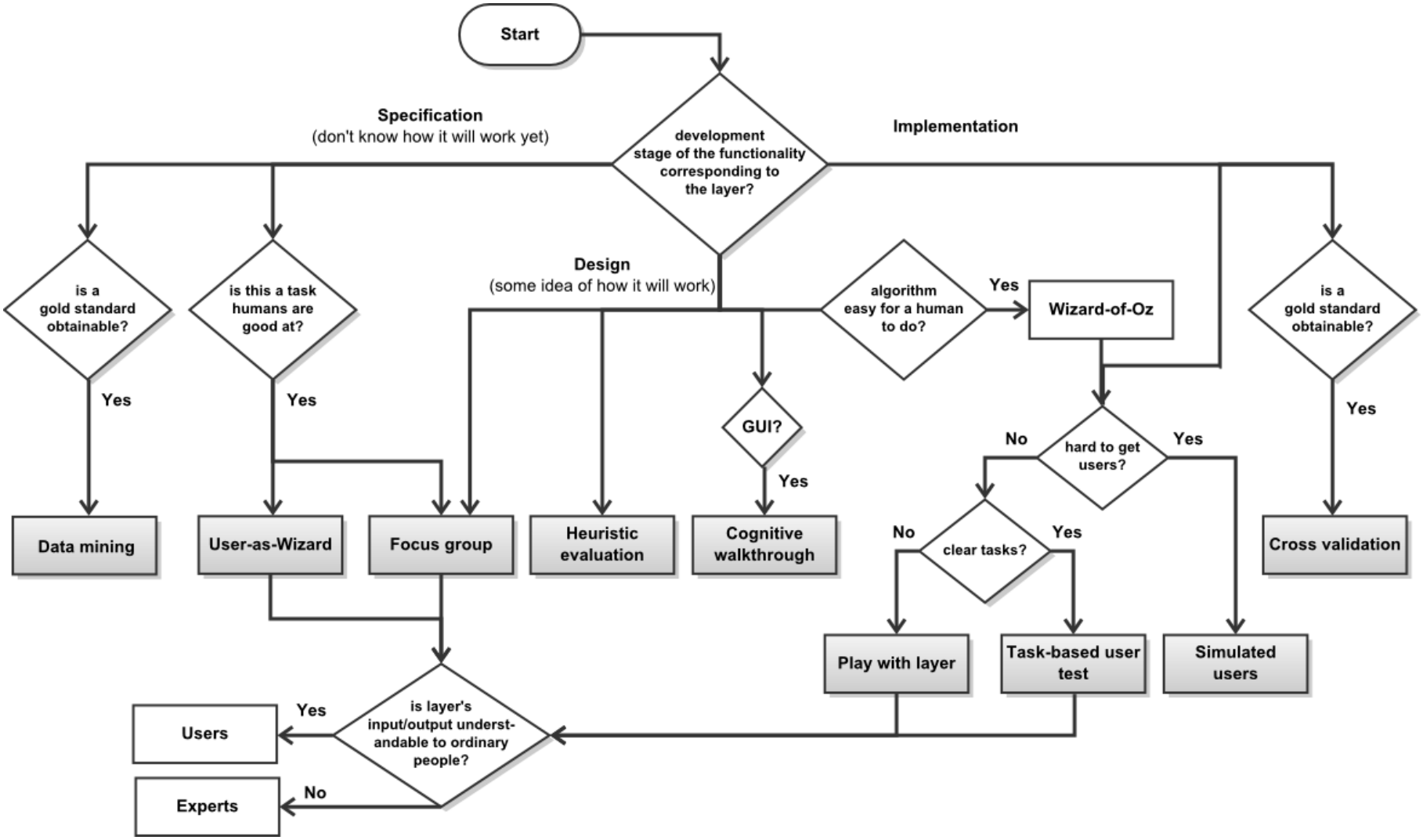

Figure 9: **A partial organization of rules of thumb into a decision process for the selection of methods to use in the layered evaluation of IAS**

6. Limitations and Alternative Approaches

In this section we discuss the framework's scope, focusing on areas that restrict its applicability. Following that, we provide a brief account the potential limitations of formative evaluation, and how the discussion of formative evaluation methods is also relevant to summative evaluation. The section closes with a discussion of complementary and alternative frameworks and approaches that have been proposed for the evaluation of IAS, and that one may want to consider in order to alleviate some of the discussed shortcomings of the proposed approach.

6.1. Scope and Limitations of the Proposed Framework

The proposed framework is intended to be applicable to as wide a range of IAS as possible, independently of their application domain, type and purpose of adaptation, etc. It is meant to guide the design at different stages of the development lifecycle of an adaptive system. The framework itself is intentionally not prescriptive in terms of evaluation methods, techniques, and data collection approaches, but strives to provide guidance for evaluators to make informed decisions on these matters. Although it can be readily used to inform the design of summative studies of specific aspects of an adaptive system's behaviour, it has been primarily conceived to facilitate the planning and undertaking of principled formative studies.
The framework does exhibit a number of limitations that should be taken into account when applying it. These relate primarily to the applicability of layers in certain types of IAS, aspects of adaptation not directly addressed by the framework, and, arguably, the feasibility (in terms of temporal and resource constraints) of applying the framework in its entirety.

It has already been discussed that some of the proposed layers may not be possible to evaluate separately in a system, or, for that matter, may not even exist – a fact probably obvious for the case of the first and last layers in the framework, but not exclusive to them. For instance, for systems that use inference mechanisms which relate input data and adaptation decisions directly –as sometimes found in machine learning systems (Krogsæter, Oppermann, & Thomas, 1994; Pohl, 1997, 1999)–, the modelling layer might not be applicable in isolation. One possible way of mitigating this type of problems, namely the combination of layers so that the resulting adaptation process stages (and corresponding layers) better reflect the system's actual operation, has been discussed in detail in Section 5.2.

At a different level, the framework deliberately does not address the evaluation of meta-adaptivity. The term is used here to refer to systems that are capable of assessing and modifying their own adaptive behaviour, learning, in the process, to identify situations in which different adaptations are best applied. Although there are different forms and levels of sophistication of meta-adaptivity, some of which not even computationally possible yet (Totterdell & Rautenbach, 1990), all of these have one characteristic in common: they require that a system be capable of self-evaluating its own adaptive behaviour. In more detail, this refers to the run-time assessment of the effects of decided upon and effected adaptations, with the intent of evaluating their "success" (i.e., whether the goals underlying their introduction have been met). This stage is referred to as "second-level adaptation" in Totterdell & Rautenbach (1990) and may further involve the modification of aspects of the lower-level adaptation cycle (e.g., by enabling or disabling rules in rule-based adaptation, or by altering the "weight" of alternatives, in decision theory-based adaptation).

The evaluation of meta-adaptivity is, as one might expect, a complicated matter. Practically, it necessitates the consideration of an additional second-level adaptation process, comprising: identification/isolation of the effects of applied adaptations on the user's behaviour; comparison between said effects and the ones intended or desired; and, potentially, selection and application of alternative sets of behaviours. A plausible evaluation approach

may involve ensuring that the system shares the same views as the users with regards to the "success", or "failure" of adaptations. Seen from a different perspective, if an IAS assesses and modifies lower-level adaptation "strategies", then what needs to be evaluated is whether any such modifications are optimal from the perspective of the user. Although, from an engineering standpoint, the IAS component(s) involved in "adapting the adapter" operate at a meta-level with respect to the rest of the IAS components, this distinction may not be relevant from the perspective of evaluation. For certain systems it may be possible, for example, to treat "meta-adaptations" as just another type of adaptation. This would mean that meta-level adaptations are amenable to the same treatment as first-level ones, and can thus be included in the layered evaluation as this has been described so far. To the best of our knowledge, there do not yet exist proposals in the literature for generically addressing this challenging topic.

Another limitation lies in the breadth of the framework. Applying all layers and criteria to a single system, potentially at various stages of the development process, is next to impossible. As mentioned earlier, the framework is meant to inform and guide study design decisions. It may neither be feasible nor necessary to apply all layers and criteria. For example, the *Collection of Input Data* layer has not been addressed in evaluation studies of many systems. If there are no obvious shortcomings in this layer, the evaluation of other layers may take priority. Nevertheless (and this is an important implication of the layered approach), due to the implicit dependencies of layers, evaluators need to be aware that a problem identified in a higher layer might just be the symptom of problems introduced at lower layers.

The above are some of the limitations of the framework's scope, but not necessarily the only ones. We fully expect that there will exist evaluation settings and system features that may render the framework inapplicable. We encourage evaluators to critically consider the framework in those cases, and, where applicable, modify and extend it to fit their needs.

6.2. Limitations of Formative Evaluation

This paper has focused almost exclusively on formative methods. This is not to be interpreted as a preference or indeed as an implicit suggestion of a superiority of formative evaluation. The relative merits of formative versus summative evaluation have been hotly debated (e.g. Cronbach et al., 1980; Scriven 1981, 1991; Chen, 1996). Some of the limitations mentioned for formative evaluation are that:

- Formative evaluations may be more time- and labour- intensive compared to most forms of summative evaluation due to relying more on qualitative methods.
- Formative evaluations do not seek to generalize, so may be more limited in their findings.
- Formative evaluations are not necessarily as carefully controlled; they are typically not aimed at producing scientific proof.
- Formative evaluations may be less suitable for comparisons, as they do not necessarily produce an objective measure of "goodness".
- Formative evaluations may be less independent, with more involvement of the design team.

These limitations do not necessarily always hold: they depend on how the formative studies are set up. Additionally, methods are not necessarily either formative or summative in nature; a single study may be used both to determine the system's value and how to further improve it. In fact, Scriven (1991) argued that it is a fallacy that formative and summative studies are intrinsically different. The mantra "When the cook tastes the soup, that's formative; when the guests taste the soup, that's summative" (R. Stake, as quoted in Scriven, 1991, p. 19) shows that the same method (tasting the soup) can be both formative and summative depending on when it is used and for what purpose. So, the methods presented in

this paper are not necessarily restricted to formative evaluations. Indeed, two of the methods described are arguably the most popular ones for summative evaluation of IAS (namely user-tests and cross-validation). However, the paper provided a formative perspective, for example for user tests, emphasising observational methods rather than the reporting of statistics.

Summative studies have an important role to play in demonstrating the success of the ultimate goal of the adaptive system, and their value should, thus, not be underestimated. An important role of formative studies is to produce better summative studies. For example, by improving system components through formative studies, it can be ensured that the summative studies are measuring what they are supposed to measure (e.g. how much personalization helps a student to learn) rather than being hindered by lower-level components (such as learner modelling) not working properly. Formative studies can also produce a qualitative understanding which can be verified later through well-controlled summative studies.

On a final note on the subject, summative evaluation is not restricted to evaluating the system as a whole. It is possible to perform summative evaluations using the layered evaluation framework: evaluating the "value" of individual layers. This paper has shown how formative evaluation methods can be adapted to cope with layered evaluation and the evaluation of an adaptive system. Much of this is equally applicable to summative evaluation. For example, for summative evaluation, it is just as important to ensure that the input received from the lower layers is accurate. Evaluators are urged to consider how the factors covered here may influence the design of studies and the selection of data collection instruments for summative assessment of IAS.

6.3. Complementary and Alternative Approaches

At this point it is worth briefly recounting some of the complementary, as well as alternative approaches to the evaluation of adaptation that have been proposed in the literature. Broadly speaking, some of them focus on the identification of criteria, while others address complementary aspects of adaptive systems to those of layered evaluation.

6.3.1. Identifying Appropriate Evaluation Criteria for IAS

The first framework that was designed to identify appropriate criteria was introduced by Tobar (2003). The approach is based on a so-called map which integrates different design perspectives to facilitate the understanding of adaptation assessment and design. Tobar's proposed framework is targeted towards the identification of specific adaptation features that need to be assessed, the establishment of criteria for the assessment, and the generation of evaluation plans on this basis.

A more recent approach called AnAmeter, proposed by Tarpin-Bernard, Marfisi-Schottman, and Habieb-Mammar (2009), is somewhat related to the one proposed by Tobar (2003), but has important differentiations as well. Instead of prescribing the procedural means for identifying adaptation features for assessment, AnAmeter provides a relatively exhaustive enumeration of potential adaptation constituents and determinants in an IAS in a tabular form. Evaluators can characterize the adaptivity and use the resulting table to determine exactly what needs to be assessed. This facilitates the identification of potential conflicts and correlations (e.g., where the same determinant affects several constituents). This framework is also unique in that it attempts to summarize and quantify the "degree" of adaptation in a system, and in that it is supported by a web-based tool that enables evaluators to interactively manage the tabular description of the system at hand. Although this framework is still at the early stages of its development, it appears to bear promise in structuring the adaptation space in an easy to understand way. It would also be interesting to see future work examining the extent to which this approach can be used in conjunction with layered evaluation.

6.3.2. Addressing Complementary Aspects of the Evaluation of IAS

Herder (2003) proposed a utility-based approach to complement the layered evaluation process. The basic idea is that the added value of an adaptive system can be expressed by a utility function that maps selected, measurable criteria with respect to the performance of the adaptive system to a quantitative representation. If one would compare an adaptive system with its non-adaptive counterpart, the value of adaptation is the difference in utility between the two systems. Herder (2003) argues that the main advantage of the layered evaluation approach in this context is that it separates the utility function in several functions in a principled manner.

Magoulas et al. (2003) argue about the need to develop an educational-evaluation model and a methodology that include usability testing as a standard procedure capable to determine the impact of adaptation on learners' behaviour in an educational environment. As described earlier, they introduce modifications to the standard heuristic evaluation approach and augment it with criteria that diagnose potential usability problems related to adaptation, subsequently integrating it into the layered evaluation approach. In contrast to Jameson's (2003) generic usability challenges these heuristics are formulated for the specific case of adaptive educational systems. This not only narrows their applicability but also seems to introduce some unnecessary assumptions about the system and the adaptation in particular. As shown in Section 4, Heuristic Evaluation can be used to assess several different layers, and may in particular be useful to evaluate the adaptation as a whole.

The evaluation of open learner models and their scrutability is addressed in the SMILI framework proposed by Bull and Kay (2007). As we have briefly seen, although scrutability is not itself a stage in the adaptation process, it has major implications in the evaluation of other stages, especially if users are able to modify the contents of their personal models (e.g., inaccuracies in the model may be attributable to the user's intervention, rather than to the system's derivation of incorrect assumptions). The SMILI framework allows evaluators to characterize the scrutability of a system along a set of seven different purposes of scrutability such as an increase of the user model's accuracy, or the facilitation of reflection. Different elements of the system are then rated against these purposes in order to identify useful potential evaluations, i.e., those that provide evidence of the performance of the system on one or more central purposes of the system. While the framework proposed here does address scrutability to some extent (as a criterion), the SMILI framework is by far more explicit and detailed as far as scrutability is concerned.

7. Discussion

The main postulation of layered evaluation of IAS is that adaptation needs to be decomposed and assessed in layers in order to be evaluated effectively. Since the first introduction of the term in 2000, the scientific community has adopted this concept in planning and conducting empirical studies. Many authors explicitly refer back to the foundational papers published on the topic to justify experimental designs, to provide rationale for goals or structure of their evaluation studies (Arruabarrena, Pérez, López-Cuadrado, Gutiérrez, & Vadillo, 2002; Ortigosa & Carro, 2003; Petrelli & Not, 2005; Cena et al., 2006; Goren-Bar, Graziola, Pianesi, & Zancanaro, 2006; Glahn, Specht, & Koper, 2007; Kosba, Dimitrova, & R. Boyle, 2007; Nguyen & Santos Jr, 2007; Stock et al., 2007; Carmagnola et al., 2008; Limongelli, Sciarrone, & Vaste, 2008; Ley, Kump, Maas, Maiden, & Albert, 2009; Popescu, 2009; Santos & Boticario, 2009), or to demonstrate methodological shortcomings of existing studies (Masthoff, 2002; Gena, 2005; Brusilovsky, Farzan, & Ahn, 2006; Yang & Huo, 2008; Brown, Brailsford, Fisher, & Moore, 2009). The fact that layered evaluation received such a high level of attention in the literature reaffirms the claim that the evaluation of adaptive systems implicates some inherent difficulties.

The benefits of layered evaluation are perhaps representatively illustrated by a set of studies of a mobile adaptive multimedia guide system for museums called PEACH (Stock & Zancanaro, 2007). PEACH records the visitors' movements through the museum and collects explicit feedback about items seen. Based on this data, PEACH provides recommendations and a personalised report presented through a life-like agent. PEACH has been evaluated in a number of different empirical studies involving the running system respectively prototypes of the system. The studies can be associated with different evaluation layers.

In regard to *data collection*, the user can express preferences through a so-called "like-o-meter". A field study with 140 users showed that visitors are willing to provide their feedback. They understood how to use the feedback system and provided a sufficient number of ratings (Stock et al., 2007). The study provided evidence that the tool is effective in collecting feedback from visitors.

In regard to the *modelling* of users, the movements of visitors in the museum were recorded and categorized into different behaviour patterns (Zancanaro, Kuflik, Boger, Goren-Bar, & Goldwasser, 2007). Clustering algorithms confirmed existing qualitative ethnographic findings on visitor behaviour.

In regard to the *adaptation decision*, a study was designed to explore which "adaptivity dimensions" would be accepted by users, i.e., are presentations that rely on one characteristic in the user model preferred over decisions that rely on different characteristics? In a laboratory study, users were presented with two simulated systems, one being adaptive and the other non-adaptive. After expressing their preference for one of the versions, they were asked to give reasons for their preference in regard to the four dimensions the system can adapt to (location, interest, follow-up, history). The study yielded insights with respect to the dimensions of adaptation which may be accepted by different user groups (Goren-Bar et al., 2006).

In regard to the *instantiation of adaptation* it should be noted that the user interface of the museum guide evolved over several user-centred design cycles (Goren-Bar et al., 2005). One of the interface components on the mobile device is a life-like character that presents information and engages the user. The effectiveness of this character in attracting the visitors' attention was tested in a study with an early prototype (Kruppa & Aslan, 2005).

The combination of these studies of PEACH and the improvements made based on their results contributed to the successful deployment of a full adaptive system in a real-world environment.

The evaluation framework proposed here is centred around a decomposition model that identifies five distinct stages in the adaptation process that should be evaluated as individual (or combined) layers. An important point we would like to make about the proposed decomposition is that it is neither the only one feasible, nor, necessarily, the most appropriate one for all types of assessment of IAS one might want to perform. For instance, it would be possible to decompose adaptation on the basis of the software components involved in a system's implementation. Furthermore, even if one takes a process-based approach to the decomposition, it is not necessary that the same level of granularity be employed. Our proposal tries to strike a balance between, on the one hand, identifying all the individual clusters of steps involved in that process, and, on the other hand, having a manageable set of coherent and assessable "targets". A related point that merits attention is that evaluating an IAS in a layered fashion (irrespectively of whether the proposed model is followed), does not directly address "cross-cutting" evaluation concerns, which have implications on all adaptation stages. Evaluators are still required to ensure that such concerns are individually integrated into the evaluation activities of each stage.

The proposed framework's target audience includes potentially most of the actors involved in the development of adaptive systems (e.g., usability experts, system designers, evaluators), as the framework may be employed from different perspectives. While a

practitioner might use it to improve an existing system, a researcher might apply the framework to several systems in order to compare the quality of different modelling or inference approaches for the same task. The framework thus aims to serve both as an instrument to be used for the principled design of evaluation studies of IAS, but also as the common ground between disciplines for the derivation of concrete, validated design knowledge for different types of adaptation in a variety of application domains.

Another goal of the framework is to facilitate the integration of evaluation activities in the iterative design of IAS. Evaluation can (and, arguably, should) take place throughout a system's development, from early on to inspire the design of adaptive behaviour, up until and including the implementation and deployment of a system. In this context, the results of formative evaluation can be quite important in terms of system evolution: most often, evaluations are not just intended to investigate how good a layer (or system) is, but seek insights over what causes problems and why. On the other hand, summative evaluation of either individual layers or a system as a whole are also of paramount importance, as they offer a solid basis for generalization of findings, and foster theory development, which has been a perennial goal of the IAS field.

Skill is required in isolating and evaluating (combinations of) layers in a system's adaptive behaviour. We have shown examples of how this can be done, several of them grounded on evaluation work reported in the literature.

Normally, multiple evaluation methods will be used during the development of an IAS. Adaptive systems can clearly benefit from the many methods available in the field of HCI, to involve users in system design and evaluation. This paper has shown how these traditional methods need to be tailored to suit the particular requirements of adaptivity in the user-system interaction. It has also described some methods (e.g., User-as-Wizard) that are specific to the adaptive systems field. The best method to employ at any one time will primarily depend on when the evaluation takes place (with respect to the system's development lifecycle) and the characteristics of the layer under consideration. We have addressed this topic, as well as two other areas where the application of the framework necessitates that domain- and system-specific characteristics be taken into account: (a) the potential combination of layers for the purposes of the evaluation, and (b) the derivation of metrics/criteria that reflect the goals of employing adaptation in a system in the first place, and can be used to assess its performance at different stages of the adaptation cycle. Our express aim in pursuing the above goals has been to remain non-prescriptive, yet provide a sufficiently holistic approach so that it can be readily employed in the evaluation of IAS.

In closing, the concepts behind layered evaluation have already had a significant impact on the evaluation of IAS. It is our hope that this paper will foster the wider adoption of the approach and will contribute to an increase in the number and quality of studies in the field.

Acknowledgements

We would like to thank the anonymous reviewers and the UMUAI editor who have helped us to substantially improve the quality of the paper. We would also like to thank the participants of the several workshops and tutorials that we have organized on the subject of the evaluation of adaptive systems for their constructive input to the evolution of the work presented here.

References

Ackerman, M. S., Cranor, L. F., & Reagle, J. (1999). Privacy in e-commerce: examining user scenarios and privacy preferences. In: *1st ACM Conference on Electronic Commerce*, Denver, CO: ACM, pp. 1-8.

Albrecht, D., & Zuckerman, I. (Eds.) (2007). Statistical and probabilistic methods for user

modeling. Special Issue, *User Modeling and User Adapted Interaction, 17*(1-2), 1-215.
Anderson, J. R., Boyle, C. F., & Yost, G. (1985). The geometry tutor. In: *9th International Joint Conference on Artificial Intelligence* (Los Angeles, CA), San Francisco: Morgan Kaufmann, pp. 1-7.
Arruabarrena, R., Pérez, T., López-Cu adrado, J., Gutiérrez, J., & Vadillo, J. (2002). On evaluating adaptive systems for education. In: *2nd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (Malaga, Spain), LNCS 2347, Berlin: Springer, pp. 363-367.
van Barneveld, J. and van Setten, M. (2003). Involving users in the design of user interfaces for TV recommender systems. In: *3rd Workshop on Personalization in Future TV* at UM03, Johnstown, PA.
Batliner, A., Steidl, S., Hacker, C., & Nöth, E. (2008). Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech, *User Modeling and User Adapted Interaction*, *18*(1-2), 175–206.
Berkovsky, S., Kuflik, T., & Ricci, F. (2008). Mediation of user models for enhanced personalization in recommender systems. *User Modeling and User-Adapted Interaction*, *18*(3), 245–286.
Beyer, H. & Holtzblatt, K. (1997). *Contextual design: Defining customer-centred systems*. San Francisco: Morgan Kaufmann.
Billsus, D., & Pazzani, M. (1999). A hybrid user model for news story classification. In: *7th International Conference on User Modeling* (Banff, Canada), Vienna: Springer, pp. 98-108.
Bontcheva, K., & Wilks, Y. (2005). Tailoring automatically generated hypertext. *User Modeling and User-Adapted Interaction*, *15*(1-2), 135-168.
Boyle, C., & Encarnacion, A. O. (1994). MetaDoc: An adaptive hypertext reading system. *User Modeling and User-Adapted Interaction*, *4*(1), 1-19.
Brown, E. J., Brailsford, T. J., Fisher, T., & Moore, A. (2009). Evaluating learning style personalization in adaptive systems: Quantitative methods and approaches. *IEEE Transactions on Learning Technologies*, *2*(1), 10-22.
Browne, D., Norman, M., & Riches, D. (1990). Why build adaptive systems. In D. Browne, P. Totterdell, & M. Norman (Eds.), *Adaptive user interfaces*. London: Academic Press, pp. 15-57.
Brusilovsky, P., & Eklund, J. (1998). A study of user model based link annotation in educational hypermedia. *Journal of Universal Computer Science*, *4*(4), 429–448.
Brusilovsky, P., Karagiannidis, C., & Sampson, D. (2001). The benefits of layered evaluation of adaptive applications and services. In: *1st Workshop on Empirical Evaluation of Adaptive Systems* at UM2001*,* Sonthofen, Germany, pp. 1–8.
Brusilovsky, P., Farzan, R., & Ahn, J. (2006). Layered evaluation of adaptive search. In: *Workshop on Evaluating Exploratory Search Systems* at SIGIR06, Seattle, WA, pp. 11-13.
Bull, S., & Kay, J. (2007). Student models that invite the learner in: The SMILI☺ open learner modelling framework. *International Journal of Artificial Intelligence in Education*, *17*(2), 89-120.
de Campos, L., Fernández-Luna, J., Huete, J., & Rueda-Morales, M. (2009). Managing uncertainty in group recommending processes. *User Modeling and User-Adapted Interaction*, *19*(3), 207-242.
Card, S.K., Thomas, T.P., & Newell, A. (1983). *The Psychology of Human-Computer Interaction*, London: Lawrence Erbaum Associates.
Carmagnola, F., Cena, F., Console, L., Cortassa, O., Gena, C., Goy, A., Torre, I., Toso, A., & Vernero, F. (2008). Tag-based user modeling for social multi-device adaptive guides. *User Modeling and User-Adapted Interaction*, *18*(5), 497-538.

Carmichael, D. J., Kay, J., & Kummerfeld, B. (2005). Consistent modelling of users, devices and sensors in a ubiquitous computing environment. *User Modeling and User-Adapted Interaction*, *15*(3-4), 197-234.

Carroll, J. M. (2000). Five reasons for scenario-based design. *Interacting with Computers*, *13*(1), 43-60.

Cena, F., Console, L., Gena, C., Goy, A., Levi, G., Modeo, S., & Torre, I. (2006). Integrating heterogeneous adaptation techniques to build a flexible and usable mobile tourist guide. *AI Communications*, *19*(4), 369-384.

Chen, H.T. (1996), A comprehensive typology for program evaluation. *American Journal of Evaluation, 17*(2), 121-130.

Cheverst, K., Byun, H. E., Fitton, D., Sas, C., Kray, C., & Villar, N. (2005). Exploring issues of user model transparency and proactive behaviour in an office environment control system. *User Modeling and User-Adapted Interaction*, *15*(3-4), 235-273.

Chickering, D.M, & Paek, T. (2007). Personalizing influence diagrams: applying online learning strategies to dialogue management. *User Modeling and User Adapted Interaction*, *17*(1-2), 71–91.

Chin, D. (2001). Empirical evaluation of user models and user-adapted systems. *User Modeling and User-Adapted Interaction*, *11*(1-2), 181-194.

Claypool, M., Le, P., Wased, M., & Brown, D. (2001). Implicit interest indicators. In: *International Conference on Intelligent User Interfaces*, Santa Fe, NM: ACM, pp. 33-40.

Conati, C., & MacLaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User Adapted Interaction*, *19*(3), 267–303.

Cooper, A. (1999). *The inmates are running the asylum*. Indianapolis: Macmillan.

Cooper, D., Arroyo, I., Woolf, B., Muldner, K., Burleson, W., & Christopherson, R. (2009). Sensors model student self concept in the classroom. In: *1st International Conference* on *User Modeling, Adaptation, and Personalization* (Trento, Italy), LNCS 5535, Berlin: Springer, pp. 30-41.

Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., Aroyo, L., & Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, *18*(5), 455-496.

Cronbach, L.J., Ambron, S.R., Dornbusch, S.M., Hess, R.D., Hornik, R.C., Philips, D.C., Walker, D.F., & Weiner, S.S. (1980). *Toward reform of program evaluation*. San Francisco, CA: Jossey-Bass.

Czarkowski, M. (2006) *A scrutable adaptive hypertext*. PhD Thesis, University of Sydney.

Damiano, R., Gena, C., Lombardo, V., Nunnari, F., & Pizzo, A. (2008). A stroll with Carletto: adaptation in drama-based tours with virtual characters. *User Modeling and User Adapted Interaction*, *18*(5), 417–453.

De Bra, P., Houben, G., & Wu, H. (1999). AHAM: a Dexter-based reference model for adaptive hypermedia. In: *10th ACM Conference on Hypertext and Hypermedia,* Darmstadt, Germany: ACM, pp. 147-156.

Degemmis, M., Lops, P., & Semeraro, G. (2007). A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation. *User Modeling and User-Adapted Interaction*, *17*(3), 217-255.

Dey, A. K., & Abowd, G. D. (2000). Towards a better understanding of context and context-awareness. In: *Workshop on the What, Who, Where, When, and How of Context-Awareness* at CHI 2000, The Hague, The Netherlands, pp. 304–307.

Dix, A., Finlay, J., Abowd, G., & Beale, R. (1998) *Human Computer Interaction*, 2nd Ed. Englewood Cliffs, NJ: Prentice-Hall.

D'Mello, S.K., Craig, S.D., Sullins, J., & Graesser, A.C. (2006). Predicting affective states through an emote-aloud procedure from AutoTutor's mixed-initiative dialogue. *International Journal of Artificial Intelligence in Education*, *16*(1), 3–28.

D'Mello, S.K., Craig, S.D., Witherspoon,A., McDaniel, B., & Graesser, A.C. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*, *18*(1-2), 45-80.

Domshlak, C., & Joachims, T. (2007). Efficient and non-parametric reasoning over user preferences. *User Modeling and User-Adapted Interaction*, *17*(1-2), 41-69.

Dumas, J.S. and Loring, B.A. (2008). *Moderating usability tests: principles and practices for interacting*. San Francisco: Morgan Kaufmann.

Encarnação, L., & Stoev, S. (1999). Application-independent intelligent user support system exploiting action-sequence based user modeling. *7th International Conference on User Modeling* (Banff, Canada), Vienna: Springer, pp. 245-254.

Ericsson, K.A., & Simon, H.A. (1993). *Protocol analysis: Verbal reports as data. Revised edition*. Cambridge, MA: MIT Press.

Forbes-Riley, K., Rotaru, M., & Litman, D.J. (2008). The relative impact of student affect on performance models in a spoken dialogue tutoring system. *User Modeling and User-Adapted Interaction*, *18*(1-2), 11-43.

Gena, C. (2005). Methods and techniques for the evaluation of user-adaptive systems. *The Knowledge Engineering Review*, *20*(1), 1-37.

Gena, C., & Ardissono, L., (2004). Intelligent support to the retrieval of information about hydric resources. *3rd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, LNCS 3137, Berlin: Springer, pp. 126-135.

Gena, C., & Weibelzahl, S. (2007). Usability engineering for the adaptive web. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web: Methods and strategies of web personalization,* Berlin: Springer, pp. 720-762.

George, S., Zukerman, I., & Niemann, M. (2007). Inferences, suppositions and explanatory extensions in argument interpretation. *User Modeling and User Adapted Interaction*, *17*(5), 439-474.

Glahn, C., Specht, M., & Koper, R. (2007). Smart indicators on learning interactions. In: *2nd European Conference on Technology Enhanced Learning* (Crete, Greece), LNCS 4753, Berlin: Springer, pp. 56-70.

Goecks, J., & Shavlik, J. (2000). Learning users' interests by unobtrusively observing their normal behavior. In: *5th International Conference on Intelligent User Interfaces*, New Orleans, LA: ACM, pp. 129-132.

Goren-Bar, D., Graziola, I., Pianesi, F., & Zancanaro, M. (2006). The influence of personality factors on visitor attitudes towards adaptivity dimensions for mobile museum guides. *User Modeling and User-Adapted Interaction*, *16*(1), 31-62.

Goren-Bar, D., Graziola, I., Rocchi, C., Pianesi, F., Stock, O., & Zancanaro, M. (2005). Designing and redesigning an affective interface for an adaptive museum guide. In: *1st International Conference on Affective Computing and Intelligent Interaction* (Beijing, China), LNCS 3784, Berlin: Springer, pp. 939-946.

Gould, J., Conti, J., & Hovanyecz, T. (1982). Composing letters with a simulated listening typewriter. In: *1st ACM Conference on Human Factors in Computer Systems (CHI),* Gaithersburg, MD: ACM, pp. 367-370.

Gould, J. D., & Lewis, C. (1985). Designing for usability: key principles and what designers think. *Communications of the ACM*, *28*(3), 300-311.

Green, D. P., Ha, S. E., & Bullock, J. G. (2010). Enough already about "black box" experiments: Studying mediation is more difficult than most scholars suppose. *ANNALS of the American Academy of Political and Social Science, 628*(1), 200-208.

Grudin, J., & Pruitt, J. (2002). Personas, participatory design, and product development: An infrastructure for engagement. In: *Participatory Design Conference*, Malmö, Sweden: ACM, pp. 144-161.

Guzmán, E., Conejo, R., & Pérez-de-la-Cruz, J.-L. (2007). Adaptive testing for hierarchical

student models. *User Modeling and User-Adapted Interaction, 17*(1-2), 119–157.
van den Haak, M.J., de Jong, M.D.T., & Schellens, P.J. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology, 22*(5), 339-351.
van den Haak, M.J., de Jong, M.D.T., & Schellens, P.J. (2004). Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison. *Interacting with Computers, 16*(6), 1153-1170.
Herder, E. (2003). Utility-based evaluation of adaptive systems. In: *2nd Workshop on Empirical Evaluation of Adaptive Systems* at UM2003, Johnstown, PA, USA. pp. 25-30.
Herlocker, J.L., Konstan, J.A., Terveen, L.G., & Riedl, J.T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems, 22*(1), 5–53.
Hertzum, M., Hansen, K., Andersen, H.H.K. (2009). Scrutinizing usability evaluation: Does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology, 28*(2), 165-181.
Hollink, V., van Someren, M., & Wielinga, B. (2007). Discovering stages in web navigation for problem-oriented navigation support. *User Modeling and User Adapted Interaction*, *17*(1-2), 183–214.
Höök, K. (2000). Steps to take before intelligent user interfaces become real. *Interacting with Computers*, *12*(4), 409-426.
Hoppe, H., Tauber, M., & Ziegler, J. (1986). A survey of models and formal description methods in HCI with example applications. *ESPRIT Project*, *385*.
Horvitz, E., & Paek, T. (2007). Complementary computing: Policies for transferring callers from dialog systems to human receptionists. *User Modeling and User-Adapted Interaction*, *17*(1-2), 159-182.
Jameson, A. (2001). *Systems that adapt to their users: An integrative perspective*. Saarbrücken: Saarland University.
Jameson, A. (2003). Adaptive interfaces and agents. In: J.A. Jacko & A. Sears (Eds.), *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications*, Hillsdale, NJ: L. Erlbaum Associates, pp. 305-330.
Jameson, A. (2005). User modeling meets usability goals. In: *10th International Conference on User Modeling* (Edinburgh, UK), LNAI 3538, Berlin: Springer, pp. 1-3.
Jameson, A. (2008). Adaptive user interfaces and agents. In A. Sears & J. Jacko (Eds.), *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications*, 2nd Ed., Boca Raton, FL: CRC Press, pp. 433-458.
Jameson, A. (2009). Understanding and dealing with usability side effects of intelligent processing. *AI Magazine*, *30*(4), 23-40.
Jameson, A., & Schwarzkopf, E. (2002). Pros and cons of controllability: An empirical study. In: *2nd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (Málaga, Spain), LNCS 2347, Berlin: Springer, pp. 193-202.
Kaplan, C., Fenwick, J., & Chen, J. (1993). Adaptive hypertext navigation based on user goals and context. *User Modeling and User-Adapted Interaction, 3*(3), 193-220.
Karagiannidis, C., & Sampson, D. (2000). Layered evaluation of adaptive applications and services. In: *1st International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (Trento, Italy), LNCS 1892, Berlin: Springer, pp. 343-346.
Kay, J. (2000). Stereotypes, student models and scrutability. *5th International Conference on Intelligent Tutoring Systems* (Montréal, Canada), LNCS 1839, Berlin: Springer, pp. 19-30.
Kay, J. (2001). Learner control. *User Modeling and User-Adapted Interaction*, *11*(1-2), 111-127.
Knutov, E., De Bra, P., & Pechenizkiy, M. (2009). AH - 12 years later: a comprehensive survey of adaptive hypermedia methods and techniques. *New Review of Hypermedia and*

Multimedia, *15*(1), 5-38.
Kobsa, A. (2007). Privacy-enhanced web personalization. In: P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web: Methods and strategies of web personalization,* Berlin: Springer, pp. 628-670.
Koch, N., & Wirsing, M. (2002). The Munich reference model for adaptive hypermedia applications. In: *2nd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (Málaga, Spain), LNCS 2347, Berlin: Springer, pp. 213-222.
Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *14th International Joint Conference on Artificial Intelligence* (Montréal, Canada), San Francisco: Morgan Kaufmann, pp. 1137-1145.
Kosba, E., Dimitrova, V., & Boyle, R. (2007). Adaptive feedback generation to support teachers in web-based distance education. *User Modeling and User-Adapted Interaction*, *17*(4), 379-413.
Krogsæter, M., Oppermann, R., & Thomas, C. G. (1994). A user interface integrating adaptability and adaptivity. In: R. Oppermann (Ed.), *Adaptive user support: ergonomic design of manually and automatically adaptable software*, Hillsdale, NJ: Lawrence Erlbaum, pp. 97-125.
Krueger, R., & Casey, M. (2009). *Focus groups: A practical guide for applied research*. 4th Ed., Los Angeles: Sage Publications.
Kruppa, M., & Aslan, I. (2005). Parallel presentations for heterogeneous user groups – an initial user study. In: *4th International Conference on Intelligent Technologies for Interactive Entertainment (Madonna di Campiglio, Italy)*, LNAI 3814, Berlin: Springer, pp. 54-63.
Law A., Freer, Y., Hunter, J., Logie, R., McIntosh, N., & Quinn, J. (2005). A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit, *Journal of Clinical Monitoring and Computing*, *19*(3), 183-194.
Lekakos, G., & Giaglis, G. (2007). A hybrid approach for improving predictive accuracy of collaborative filtering algorithms. *User Modeling and User-Adapted Interaction*, *17*(1-2), 5-40.
Lewis C. (1982). Using the 'thinking-aloud' method in cognitive interface design. *Research report RC9265*; IBM T.J. Watson Research Center, Yorktown Heights, NY
Ley, T., Kump, B., Maas, A., Maiden, N., & Albert, D. (2009). Evaluating the adaptation of a learning system before the prototype is ready: A paper-based lab study. In: *1st International Conference on User Modeling, Adaptation, and Personalization* (Trento, Italy), LNCS 5535, Berlin: Springer, pp. 331-336.
Limongelli, C., Sciarrone, F., & Vaste, G. (2008). LS-Plan : An effective combination of dynamic courseware generation and learning styles in web-based education. In: *5th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (Hannover, Germany), LNCS 5149, Berlin: Springer, pp. 133-142.
MacLaren, B., & Koedinger, K. (2002). When and why does mastery learning work: Instructional experiments with ACT-R "SimStudents". In: *6th International Conference on Intelligent Tutoring Systems*, (Biarritz, France), Berlin: Springer, pp. 355-366.
Magoulas, G. D., Chen, S. Y., & Papanikolaou, K. A. (2003). Integrating layered and heuristic evaluation for adaptive learning environments. In: *2nd Workshop on Empirical Evaluation of Adaptive Systems* at UM2003, Johnstown, PA, pp. 5-14.
Maguire, M. (2001). Methods to support human-centred design. *International Journal Human-Computer Studies*, *55*(4), 587- 634.
Masthoff, J. (2002). The evaluation of adaptive systems. In: N. Patel (Ed.), *Adaptive Evolutionary Information Systems*, London: Idea Group Publishing, pp. 329-347.
Masthoff, J. (2004). Group modeling: Selecting a sequence of television items to suit a group

of viewers. *User Modeling and User Adapted Interaction*, *14*(1), 37-85.
Masthoff, J. (2006). The user as wizard: A method for early involvement in the design and evaluation of adaptive systems. In: *5th Workshop on User-Centred Design and Evaluation of Adaptive Systems* at AH06, Dublin, Ireland, pp. 460-469.
Masthoff, J. (unpublished). *Automatically constructing good hierarchies: HCI meets AI*.
Masthoff, J., & Gatt, A. (2006). In pursuit of satisfaction and the prevention of embarrassment: Affective state in group recommender systems. *User Modeling and User-Adapted Interaction*, *16*(3-4), 281-319.
Masthoff, J., Vasconcelos, W.W., Aitken, C., & Correa da Silva, F.S. (2007). Agent-based group modelling for ambient intelligence, *AISB Symposium on Affective Smart Environments*, Newcastle, UK.
Maulsby, D., Greenberg, S., & Mander, R. (1993) Prototyping an intelligent agent through wizard of Oz. In: *10th ACM Conference on Human Factors in Computing Systems*, Amsterdam, The Netherlands: ACM, pp. 277-284.
McNee, S.M., Riedl, J., & Konstan, J.A. (2006). Being accurate is not enough: How accuracy metrics have hurt recommender systems. In: *CHI work in progress*, Montréal, Canada.
Miettinen, M., & Oulasvirta, A. (2007). Predicting time-sharing in mobile interaction, *User Modeling and User Adapted Interaction*, *17*(5), 475–510.
Millán, E., & Pérez de la Cruz, J.L. (2002). Diagnosis algorithm for student modeling diagnosis and its evaluation. *User Modeling and User Adapted Interaction*, *12*(2–3), 281–330.
Mobasher, B. (2007). Data mining for web personalization. In: P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web: Methods and strategies of web personalization*, Berlin: Springer, pp. 90-135.
Mobasher, B., & Tuzhilin, A. (Eds.) (2009). Special issue on data mining for personalization. *User Modeling and User Adapted Interaction,* 19 (1-2), 1-166.
Moncur, W., Masthoff, J., & Reiter, E. (2008). What do you want to know? Investigating the information requirements of patient supporters. In: *21th IEEE International Symposium on Computer-Based Medical Systems*, Jyväskylä, Finland: IEEE, pp. 443-448.
Murray, T. (1993). Formative qualitative evaluation for 'exploratory' ITS research. *International Journal on Artificial Intelligence in Education*, 4 (2-3), 179–207.
Nguyen, H., Masthoff, J. & Edwards, P. (2007). Modelling a receiver's position to persuasive arguments. In: 2nd International Conference on Persuasive Technology (Palo Alto, CA), LNCS 4744, Berlin: Springer, pp. 271-282.
Nguyen, H., & Santos Jr, E. (2007). An evaluation of the accuracy of capturing user intent for information retrieval. In: *International Conference on Artificial Intelligence*, Las Vegas, NV: CSREA Press, pp. 341-350.
Nielsen, J. (1993). Evaluating the thinking-aloud technique for use by computer scientists. In: H.R. Hartson & D. Hix (Eds). *Advances in human-computer interaction*, Vol.3, Norwood, NJ: Ablex, pp. 69-82.
Nielsen, J. (1994a). Heuristic evaluation. In: J. Nielsen & R.L. Mack (Eds.), *Usability inspection methods*, New York: John Wiley & Sons, pp. 25-64.
Nielsen, J. (1994b). *Usability engineering,* 2nd Ed., San Francisco, CA: Morgan Kaufmann.
Norman, D. A. (1994). How might people interact with agents. *Communications of the ACM*, *37*(7), 68-71.
Nückles, M., Winter, A., Wittwer, J., Herbert, M., & Hübner, S. (2006). How do experts adapt their explanations to a layperson's knowledge in asynchronous communication? An experimental study. *User Modeling and User-Adapted Interaction*, *16*(2), 87-127.
Ogata, K. (2009). *Modern control engineering,* 5th Ed., Upper Saddle River, NJ: Prentice Hall.
Ohene-Djan, J. (2002). Ownership transfer via personalisation as a value-adding strategy for web-based education. In: *Workshop on Adaptive Systems for Web-Based Education* at

AH2002, Málaga, Spain, pp. 27-41.
Oliver, N., & Horvitz, E. (2005). A comparison of HMMs and dynamic bayesian networks for recognizing office activities. In: *10th International Conference on User Modeling* (Edinburgh, UK), LNCS 3538, Berlin: Springer, pp. 199-209.
O'Malley, C.E., Draper, S.W., & Riley, M.S. (1984). Constructive interaction: A method for studying human-computer-human interaction. In: *1st International Conference on Human-Computer Interaction*, Honolulu, HI, pp. 269-274
Oppermann, R. (1994). Adaptively supported adaptability. *International Journal of Human-Computer Studies*, *40*(3), 455-472.
Oppermann, R. (1995). Introduction. In: R. Oppermann (Ed.), *Adaptive user support: Ergonomic design of manually and automatically adaptable software*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 1-13.
Ortigosa, A., & Carro, R. M. (2003). The continuous empirical evaluation approach: Evaluating adaptive web-based courses. In: *9th International Conference on User Modeling* (Johnstown, PA), LNCS 2702, Berlin: Springer, pp. 163-167.
Paek, T., & Chickering, D.M. (2007). Improving command and control speech recognition on mobile devices: using predictive user models for language modelling. *User Modeling and User-Adapted Interaction*, *17*(1-2), 93-117.
Paramythis, A., Totter, A., & Stephanidis, C. (2001). A modular approach to the evaluation of adaptive user interfaces. In: 1st Workshop on *Empirical Evaluation of Adaptive Systems* at UM2001, Sonthofen, Germany, pp. 9–24.
Paramythis, A., & Weibelzahl, S. (2005). A decomposition model for the layered evaluation of interactive adaptive systems. In: *10th International Conference on User Modeling* (Edinburgh, UK), LNCS 3538, pp 438-442.
Person N.K., & Graesser A.C. (2002). Tutoring research group human or computer? AutoTutor in a bystander Turing test. In: *6th International Conference on Intelligent Tutoring Systems* (Biarritz, France), LNCS 2363, Berlin: Springer, pp. 821-830.
Petrelli, D., & Not, E. (2005). User-centred design of flexible hypermedia for a mobile guide: Reflections on the HyperAudio experience. *User Modeling and User-Adapted Interaction*, *15*(3-4), 303-338.
Pohl, W. (1997). LaboUr - Machine learning for user modeling. In: *7th International Conference on Human-Computer Interaction*, Amsterdam: Elsevier, pp. 27-30.
Pohl, W. (1999). Logic-based representation and reasoning for user modeling shell systems. *User Modeling and User-Adapted Interaction*, *9*(3), 217-282.
Popescu, E. (2009). Evaluating the impact of adaptation to learning styles in a web-based educational system. In: 8th International Conference on Web-Based Learning (Aachen, Germany), LNCS 5686, Berlin: Springer, pp. 343-352.
Porayska-Pomsta, K., Mavrikis, M., & Pain, H. (2008). Diagnosing and acting on student affect: the tutor's perspective. *User Modeling and User Adapted Interaction*, *18*(1-2), 125-173.
Preece, J., Rogers, Y., Sharp, H., & Benyon, D. (1994). *Human–Computer Interaction*. Reading, MA: Addison-Wesley.
Robson, C. (1994). *Experiment, design, and statistics in psychology*. 3rd Ed. London: Penguin.
de Rosis, F., Mazzotta, I., Miceli, M., & Poggi, I. (2006). Persuasion artifices to promote wellbeing. In: *1st International Conference on Persuasive Technology* (Eindhoven, The Netherlands), LNCS 3962, Berlin: Springer, pp. 84-95.
Santos, O., & Boticario, J. (2009). Guiding learners in learning management systems through recommendations. In: 4th European Conference on Technology Enhanced Learning (Nice, France), LNCS 5794, Berlin: Springer, pp. 596-601.
Schein, A.I., Popescul, A., Ungar, L.H., & Pennock, D. M. (2002). Methods and metrics for cold-start collaborative filtering. In: *25th Annual international ACM SIGIR Conference*

on Research and Development in Information Retrieval(Tampere, Finland). New York: ACM, pp. 253-260.

Schmidt, D., Zukerman, I., & Albrecht, D. (2009). Assessing the impact of measurement uncertainty on user models in spatial domains. *1st International Conference on User Modeling, Adaptation, and Personalization* (Trento, Italy), LNCS 5535, Berlin: Springer, pp. 210-222.

Scriven, M. (1981). Produce evaluation. In N.L. Smith (Ed.), *New techniques for evaluation*. Beverly Hills, CA: Sage, pp. 121-126.

Scriven, M. (1991). Beyond formative and summative evaluation. In G.W. McLaughlin and D.C. Phillips (Eds.), *Evaluation and education: At quarter century*. Chicago, IL: University of Chicago Press, pp. 19-64.

Scriven, M. (1996). Types of evaluation and types of evaluator. *Evaluation Practice, 17*(2), 151-162.

Serna, A. Pigot, H., & Rialle, V. (2007). Modeling the progression of Alzheimer's disease for cognitive assistance in smart homes. *User Modeling and User-Adapted Interaction*, *17*(4), 415–438.

Shneiderman, B. (1998). *Designing the user interface: Strategies for effective human-computer interaction*. Reading, MA: Addison Wesley.

Sixsmith, A. J. (2000). An evaluation of an intelligent home monitoring system. *Journal of Telemedicine and Telecare*, *6*(2), 63-72.

Spada, D., Sánchez-Montañés, M., Paredes, P., & Carro, R. (2008). Towards inferring sequential-global dimension of learning styles from mouse movement patterns. In: *5th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (Hannover, Germany), LNCS 5149, Berlin: Springer, pp. 337-340.

Stary, C., & Totter, A. (1997). How to integrate concepts of the design and the evaluation of adaptable and adaptive user interfaces. 3rd ERCIM Workshop on User Interfaces for All, Obernai, France, pp. 68-75.

Stamou, S., & Ntoulas, A. (2009). Search personalization through query and page topical analysis. *User Modeling and User-Adapted Interaction*, *19*(1), 5-33.

Stock, O., & Zancanaro, M. (2007). *PEACH - Intelligent interfaces for museum visits*. Berlin: Springer.

Stock, O., Zancanaro, M., Busetta, P., Callaway, C., Krüger, A., Kruppa, M., Kuflik, T., et al. (2007). Adaptive, intelligent presentation of information for the museum visitor in PEACH. *User Modeling and User-Adapted Interaction*, *17*(3), 257-304.

Suebnukarn, S., & Haddawy, P. (2006). Modeling individual and collaborative problem-solving in medical problem-based learning. *User Modeling and User-Adapted Interaction*, *16*(3-4), 211-248.

Tarpin-Bernard, F., Marfisi-Schottman, I., & Habieb-Mammar, H. (2009). AnAmeter: The first steps to evaluating adaptation. *6th Workshop on User-Centred Design and Evaluation of Adaptive Systems* at UMAP2009, Trento, Italy: CEUR, pp. 11-20.

Tintarev, N., & Masthoff, J. (2008). The effectiveness of personalized movie explanations: An experiment using commercial meta-data. In: *5th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (Hannover, Germany), LNCS 5149, Berlin: Springer, pp. 204-213.

Tintarev, N., & Masthoff, J. (2007). Effective explanations of recommendations: User-centered design. In: *ACM conference on Recommender systems*, Minneapolis, MN: ACM, pp. 153-156.

Tintarev, N., & Masthoff, J. (2009). Evaluating recommender explanations: Problems experienced and lessons learned for the evaluation of adaptive systems. *6th Workshop on User Centered Design and Evaluation* at UMAP09, Trento, Italy: CEUR, pp. 54-63.

Tobar, C. M. (2003). Yet another evaluation framework. *2nd Workshop on Empirical*

Evaluation of Adaptive Systems at UM2003, Johnstown, PA, pp. 15-24.
Totterdell, P., & Boyle, E. (1990). The evaluation of adaptive systems. In D. Browne, P. Totterdell, & M. Norman (Eds.), *Adaptive user interfaces,* London: Academic Press, pp. 161-194.
Totterdell, P., & Rautenbach, P. (1990). Adaptation as a problem of design. In D. Browne, P. Totterdell, & M. Norman (Eds.), *Adaptive user interfaces,* London: Academic Press, pp. 61-84.
Totterdell, P., Rautenbach, P., Wilkinson, A., & Anderson, S. (1990). Adaptive interface techniques. In D. Browne, P. Totterdell, & M. Norman (Eds.), *Adaptive user interfaces,* London: Academic Press, pp. 131-160.
Trewin, S. (2000). Configuration agents, control and privacy. In: ACM C*onference on Universal Usability*, Arlington, VA: ACM, pp. 9-16.
Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*, 433-460.
VanLehn, K., Niu, Z., Siler, S. & Gertner, A.S. (1998). Student modeling from conventional test data: a Bayesian approach without priors. In: *5th International Conference on Intelligent Tutoring Systems* (Montréal, Canada), LNCS 1452, Berlin: Springer, pp. 434–443
van Velsen, L., van der Geest, T, Klaassen, R, Steehouder, M. (2008). User-centered evaluation of adaptive and adaptable systems: a literature review. *The Knowledge Engineering Review*, *23*(3), 261-281.
Walker, E., Rummel, N. & Koedinger, K.R, (2009). CTRL: A research framework for providing adaptive collaborative learning support. *User Modeling and User Adapted Interaction*, *19*(5), 387-431.
Wang, Y., Chen, Z., & Kobsa, A. (2006). *A collection and systematization of international privacy laws, with special consideration of internationally operating personalized websites*, http://www.ics.uci.edu/~kobsa/privacy
Weber, G., & Specht, M. (1997). User modeling and adaptive navigation support in WWW-based tutoring systems. In: *6th International Conference on User Modeling* (Chia Laguna, Italy), Vienna: Springer, pp. 289-300.
Weibelzahl, S. (2001). Evaluation of adaptive systems. In: *8th International Conference on User Modeling,* LNCS 2109, Berlin: Springer, pp. 292-294.
Weibelzahl, S. (2003). *Evaluation of adaptive systems*. PhD Thesis, University of Trier, Germany.
Weibelzahl, S. (2005). Problems and pitfalls in evaluating adaptive systems. In: *4th Workshop on the Evaluation of Adaptive Systems* at UM'05, Edinburgh, UK, pp. 57-66.
Weibelzahl, S., & Weber, G. (2003). Evaluating the inference mechanism of adaptive learning systems. In: *9th International Conference of User Modeling* (Johnstown, PA), LNCS 2702, Berlin: Springer, pp. 154-168.
Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The cognitive walkthrough method: A practitioner's guide. In: J.Nielsen.and R.L.Mack (Eds.) Usability inspection methods. New York: John Wiley & Sons, pp. 105-141.
Wilson, J., & Rosenberg, D. (1988). Rapid prototyping for user interface design. In: M. Helander (Ed.), *Handbook of human-computer interaction*, Amsterdam: Elsevier, pp. 859-875.
Winter, S., Wagner, S., & Deissenboeck, F. (2008). A comprehensive model of usability. In: *Engineering Interactive Systems Conference,* LNCS 4940, Berlin: Springer, pp. 106-122.
Witten, I.A. & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. 2nd Ed. Amsterdam: Morgan Kaufmann.
Yang, D., & Huo, H. (2008). Assessment on the adaptivity of adaptive systems. In: *International Conference on Management of e-Commerce and e-Government*, Nanchang, China: IEEE, pp. 437-440.

Yudelson, M., Medvedeva, O., & Crowley, R. (2008). A multifactor approach to student model evaluation. *User Modeling and User-Adapted Interaction*, *18*(4), 349-382.

Zancanaro, M., Kuflik, T., Boger, Z., Goren-Bar, D., & Goldwasser, D. (2007). Analyzing museum visitors' behavior patterns. In: *11th International Conference on User Modeling* (Corfu, Greece), LNCS 4511, Berlin: Springer, pp. 238-246.

Zaslow, J. (2002). If TiVo thinks you are gay, here's how to set it straight. The Wall Street Journal, sect. A, p. 1, November 26, 2002.

Ziegler, J., & Bullinger, H. J. (1991). Formal models and techniques in human-computer interaction. In: B. Shackel & S.J. Richardson (Eds.), *Human factors for informatics usability*, Cambridge, UK: Cambridge University Press, pp. 183-206.

Ziegler, C.-N., McNee, S.M., Konstan, J. A., & Lausen, G. (2005). Improving recommendation lists through topic diversification. In: *14th International World Wide Web Conference,* Chiba, Japan: ACM, pp. 22–32.

Zimmermann, A., & Lorenz, A. (2008). LISTEN: A user-adaptive audio-augmented museum guide. *User Modeling and User-Adapted Interaction*, *18*(5), 389-416.

Zimmermann, A., Specht, M., & Lorenz, A. (2005). Personalization and context management. *User Modeling and User-Adapted Interaction*, *15*(3-4), 275-302.

Zhang, T., Rau, P., & Salvendy, G. (2007). Developing instrument for handset usability evaluation: A survey study. In: *12th International Conference on Human-Computer Interaction* (Beijing, China), LNCS 4550, Berlin: Springer, pp. 662-671.

Author Biographies

Dr. Alexandros Paramythis received his Ph.D. in the area of Adaptive Systems from the Johannes Kepler University (Linz, Austria) where he is currently employed as a researcher. He has long-standing experience in the design and development of adaptive systems, gained through participation in several research projects in the field. His research interests lie in the areas of evaluation of adaptive systems, adaptive support for personalized and collaborative learning, and meta-adaptivity.

Dr. Stephan Weibelzahl is a Lecturer in the School of Computing at National College of Ireland, Dublin. He is also Principal Investigator of the National e-Learning Laboratory. Stephan has longstanding experience in the evaluation of adaptive systems. In his PhD thesis he tried to integrate current research on this topic and explored suitable evaluation methods and criteria. He has published on the problems arising in the area and adequate methods to address these problems. His research interests lie in adaptive learning systems, adaptation to motivation, user experience and blended learning.

Dr. Judith Masthoff is a Senior Lecturer in Computing Science at the University of Aberdeen. She received her Ph.D. from Eindhoven University of Technology on a thesis that described an agent-based adaptive instruction system (awarded 1997 SNS bank prize for best applied thesis of the university in that year). Her research interests lie in the areas of intelligent user interfaces, group recommender systems, persuasive technology, the evaluation of adaptive systems, personalized time-based media and automated diagrammatic reasoning. She is also involved in public engagement with science, most recently in The Joking Computer project. She currently co-leads the Computing Science discipline at the University of Aberdeen.