Analyzing Empirical Evaluation of Advanced Learning Environments: Complex Systems and Confounding Factors

Researchers of learning technologies have dedicated substantial endeavors to the design and implementation of *advanced learning environment* (ALE) which are based on various technologies at the front edge, such as multimedia and mobile communication technologies. These novel designs or implementations are given profound expectations to be influential upon learners' learning outcomes by practitioners of this field. Empirical evaluations are therefore necessary steps for appraising these systems in terms of learning outcomes.

Nevertheless, just as previous studies in the field of intelligent tutoring systems (ITS) have revealed, to evaluate the educational impact of ITS, a representative example of ALE, is a costly affair [4]. One obvious issue to be confronted with would be the inherent ambiguity and uncertainty of human learning. Besides, if the evaluation is undertaken as the form of an experiment or a quasi-experiment, *validity* and *reliability* of the quantitative evaluation are essential issues that should be soundly addressed [3]. Fortunately, ALE evaluation is neither the single nor the first case to encounter these concerns. Several long-developed research areas, including educational measurement and experimental psychology, have accumulated plenty of knowledge during past decades dealing with most of these issues.

Even so, it is still not yet the happy ending of the story. A reliable and valid evaluation of ALE remains challenging. The cause is not only due to the classic issues mentioned above, but ALEs are themselves complex artifacts in which a number of sub-components interact in a nonsimple means. It is likely that more confounding factors would be introduced into the evaluation. This scenario is a threat to classically ideal evaluation design that all variables are well controlled, except the treatment. The *complexity*, of the *evaluation task*, the *to-be-evaluated ALE itself* and the *interaction between these two systems* would be the key issue worth to be analyzed in order to dispel the fog.

We take Simon's insightful anatomy of *complex systems* as the basis for analyzing this issue [7]. The *hierarchical model* proposed by Simon to explain the complexity prevailing among natural phenomena and human-designed artifacts, though may be aged and imperfect 0, can nicely model the complexity embedded in the affair of ALE evaluation. In the rest of this article, this point of view will be further described by employing the evaluation task of adaptive educational hypermedia systems as an example for analysis.

Two Levels of Complexity: the Evaluation-wide System and the Intra-Artifact System

In the task of ALE evaluation, we identify two distinct complex systems: the *evaluation-wide system* and the *intra-artifact system*.

On the one hand, the evaluation-wide system is the abstraction of the evaluation task in accompany with various variables. These variables include the ALE to be evaluated, human participants using the ALE, and the experimental instruments such as questionnaires or psychometric tools employed to assess learners' responses. On the other hand, the intra-artifact system implicates the to-be-evaluated ALE itself such as an intelligent tutoring system for algebra, a learning management system with course sequencing rules and a museum guide system build upon handheld computing devices.

Notice that these two levels of complexity are not irrelevant to each other. There could be interactions and latent linkages between these two worlds. For example, participants of the evaluation, although are part of the evaluation-wide system, participants' usage and response are likely to be captured and modeled by some ALEs as user models to actuate the artifact's particular function. Such kind of linkage exactly reflects the complex nature of ALE evaluation.

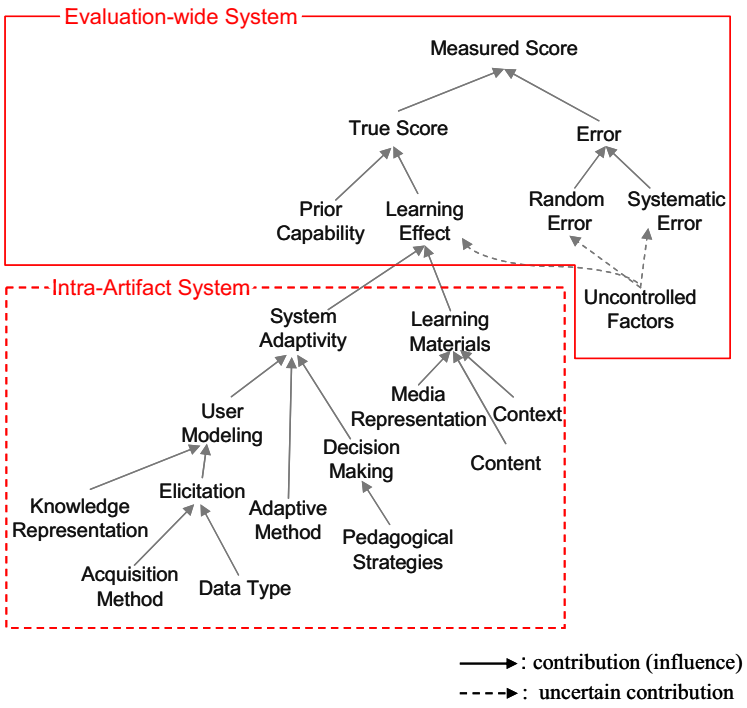

Figure 1. The hierarchy of underlying factors for AEHS

An Example of Analysis: Evaluating Adaptive Educational Hypermedia Systems

Adaptive Education Hypermedia Systems (AEHS) are user modeling-based systems that can offer learners fit-to-needs learning experience. To achieve this feature, AEHS is constituted by several sub-components which have their own distinct behaviors and properties. Figure 1 conceptually depicts the hierarchical model of this scenario. The AEHS as the intra-artifact system shown in Figure 1 can be decomposed into sub-components by considering the influence of sub-components on the performance of ones at the higher level.

On the other hand, by moving the focus to the part of evaluation-wide system, the hierarchy of the evaluation task is shown in the upper part of Figure 1. Since the evaluation affair can be treat as a kind of measurement, there is *no* so-called errorless evaluation [6]. By taking the view of classical test theory, the observed scores of learners' learning outcomes could be decomposed as:

$$S_{observed} = S_{true} + E$$

where $S_{observed}$ is the observed score derived from the measurement, S_{true} is the unknown true score and E is the measurement error. It can be observed that the construct of interest to be evaluated, learning effect, locates at the bottom of the evaluation-wide system, therefore, is not observable directly. Thus an adequate experimental design (e.g., between-groups random assignment experiment) is essential to control the influence of non-focused factors of the measurement, such as participants' prior capability and other possible covariates. It is worth noting that in Figure 1, the dashed arrows are edges representing uncertain influence. In this hierarchical model, uncontrolled factors are identified and linked to other sub-components in the system with dashed arrows.

The unique obstacle that ALE evaluation meets is the interaction and linkage between these two complex systems. Sub-components inside the intra-artifact system can also play the role of uncontrolled factors in the evaluation-wide system if the experiment is not well designed. For example, it may be possible that the user modeling modules cannot effectively capture user's information, and the AEHS may function in the way of non-adaptive manner which is less different to typical hypermedia systems. For this situation, the validity of an experimental design of with- or without- user modeling for comparative evaluation would be questionable.

Conclusions

In this article, we propose to employ hierarchical models and the conception of complex systems to realize confounding factors and their relations inside the evaluation of advanced learning environments. The complexity of the measurement and the artifact introduce more uncertainties into the evaluation affair. The approach of layered evaluation shown in [5] would be a noteworthy research direction regarding the analysis we have shown.

The analysis framework shown by this article is believed to be improvable. The theme this article disclosed, ALE evaluation as a complex system, is worth to be further investigated. The dynamics inside the complex systems 0, the concern of modern test theory [2] and the uniqueness of advance learning environments are topics to be studied as future works.

References

[1] Agre, P. E. (2003). Hierarchy and History in Simon's 'Architecture of Complexity'. *The Journal of the Learning Sciences, 12* (3), 413-426.

[2] Borsboom, D., & Mellenbergh, G. J. (2002). True Scores, Latent Variables and Constructs: A Comment on Schmidt and Hunter. *Intelligence, 30*, 505-514.

[3] Chin, D. N. (2001). Empirical Evaluation of User Models and User-Adapted Systems. *User Modeling and User-Adapted Interaction, 11*, 181-194.

[4] Iqbal, A., Oppermann, R., Patel, A., & Kinshuk (1999). A Classification of Evaluation Methods for Intelligent Tutoring Systems. In U. Arend, E. Eberleh & K. Pitschke (Eds.) *Software Ergonomie '99 - Design von Informationswelten*, Leipzig, Germany: B. G. Teubner Stuttgart, 169-181.

[5] Karagiannidis, C., Sampson, D., & Brusilovsky, P. (2001). Layered Evaluation of Adaptive and Personalized Educational Applications and Services. *Paper presented at the AIED 2001 Workshop on Assessment Methods in Web-based Learning Environments & Adaptive Hypermedia*, 21-29.

[6] Schmidt, F. L., & Hunter, J. E. (1999). Theory Testing and Measurement Error. *Intelligence, 27* (3), 183-198.

[7] Simon, H. A. (1996). *The Sciences of the Artificial* (3rd Ed.), Cambridge, MA: MIT Press.

Hao-Chuan Wang
Institute of Information Science
Academia Sinica
Taiwan
haochuan@iis.sinica.edu.tw

Tsai-Yen Li
Computer Science Department
National Chengchi University, Taiwan
li@nccu.edu.tw

Chun-Yen Chang
Department of Earth Sciences
National Taiwan Normal University
Taiwan
changcy@cc.ntnu.edu.tw