Journal of Universal Computer Science, vol. 18, no. 3 (2012), 314-326
submitted: 30/9/11, accepted: 28/1/12, appeared: 1/2/12 © J.UCS

Computer Generated Voice-Over in a Medical E-Learning Application: The Impact on Factual Learning Outcome

Stefan Minder
(University of Bern, Institute of Medical Education, Bern, Switzerland
stefan.minder@iml.unibe.ch)

Michele Notari
(Pädagogische Hochschule Bern, Bern, Switzerland
michele.notari@phbern.ch)

Felix Schmitz
(University of Bern, Institute of Medical Education, Bern, Switzerland
schmitz@iml.unibe.ch)

Rainer Hofer
(University of Bern, Institute of Medical Education, Bern, Switzerland
rainer.hofer@iml.unibe.ch)

Ulrich Woermann
(University of Bern, Institute of Medical Education, Bern, Switzerland
ulrich.woermann@iml.unibe.ch)

Abstract: The Medical Faculty of the University of Bern uses voice-over in picture driven e-learning modules to avoid split attention induced by the modality effect. To lower production costs, professional narrators have been replaced by computer-generated voices. The e-learning modules are produced with a content management system (CMS) offering text-to-speech functionality. 107 Swiss high school students passed a 20-minute e-learning sequence on cystic fibrosis. In a nested between-group design with four learning content presentation modalities (written text vs. human voice-over vs. artificial voice-over plus 15"-laptop-screens vs. 2,8"smart-phone screens), the learning outcome was assessed at three points in time: before, just after, and six weeks after the learning phase. All modalities led to significant short-term and long-term increase in factual knowledge about cystic fibrosis. Our two hypotheses are supported: (1) presenting pictures with both human and artificial voice-over leads to the same factual learning outcome, and (2) the e-learning module leads to the same learning outcome and acceptance independent of devices and their screen sizes. Furthermore, the image-voice-over modality on mobile devices (small screens) turned out to be a setting with no significant difference in effectiveness.

Keywords: text-to-speech, TTS, e-learning, modality effect, split attention
Categories: H.5.1, H.5.2, H.5.4, K.3.1

1 Introduction

The split attention principle in multimedia states: when designing instruction (including multimedia instruction) it is a goal to avoid modalities that require learners to split their attention between multiple sources of information [Mayer, 98], [Florax, 10]. A classic setting of information sources is images accompanied by written text. Schmidt-Weigand et al. [Schmidt-Weigand, 10a] state that images with text potentially foster split attention which can be avoided by substituting spoken for written text. Kürschner and Schnotz [Kürschner, 08] support the thesis that there is no difference between reading and listening comprehension when higher levels of information processing are concerned, i.e. creating mental information representation. Furthermore, when it comes to image driven e-learning on mobile devices, the limitation of screen size needs to be overcome by using customized multimedia metaphors [Nguyen, 06], e.g. there is not enough space to present pictures and corresponding text together, navigation elements are small and therefore difficult to operate, and the amount of displayable navigation elements is restricted by the screen size. Presenting images and corresponding written text in sequence mitigates the screen real estate resources problem but increases the extraneous cognitive load because of the temporal separation of the information [Schmidt-Weigand, 10a], [Paas, 10]. Again, a solution to the temporal separation problem is the replacement of written text with spoken text played simultaneously to the displayed picture. According to Cooper [Cooper, 98] cognitive load theory is based on the following principles of cognitive learning: (1) Working memory is extremely limited. (2) Long term memory is essentially unlimited. (3) The process of learning requires the working memory to be actively engaged in the comprehension (and processing) of instructional material to encode to-be-learned information into the long-term memory. (4) If the resources of the working memory are exceeded then learning will be ineffective. The integration of text-to-speech technology therefore is the most appropriate cost-, effort-, and time-effective proponent to easily incorporate the oral element of face-to-face lectures in e-learning modules [Yuanchun, 03]. Interventions redesign for learning environments integrating the capabilities of text-to-speech software will not only yield pedagogical but also financial benefits as it spares the course developer the need for high quality speech recording, re-recording, timing adjustments and other related technicalities. The interesting and novel concept brought here is not centred on the technology, as text-to-speech software is widely available and its various possible uses are obvious. However, the process of instructionally incorporating text-to-speech in the educational landscape of the university is a new pedagogical method [Rugoohputh, 09].

1.1 Spoken text in (medical) e-learning

The production of spoken text by professional narrators is a costly procedure in e-learning content production that increases the editorial complexity: while written text can easily be edited, spoken text must be re-recorded to capture content change – that is, a follow-up booking of studio and professional narrator is necessary. To avoid these additional costs in e-learning production the Medical Faculty of the University of Bern decided to switch from professional narrators to computer generated speech.

A content management system (CMS) was set up offering the production of image driven e-learning modules with fully automatic text-to-speech synthesis. Twenty e-learning modules (e.g. Larynx, pharynx, central nervous system) presenting a total of 4,320 images and more than 100,000 words of text were created within this CMS. The built-in text-to-speech software automatically produced 20 hours of voice-over that would have led to expenses of US$ 48,000 for human narrators (Swiss Association of Professional Speakers price list). Beyond the financial reasons, didactic assumptions led the management to this decision: 1) The learning outcome is better when students can listen to textual information rather than read it on screen; 2) Computer generated voices are good enough to achieve this goal; 3) With this spoken text/image combination the e-learning modules are fully functional on mobile devices with small screens achieving the same learning outcome as the standard screen versions.

1.2 Research questions and experimental requirements

To scrutinize these implications we followed the mentioned management's didactic assumptions leading to our working hypothesis as follows:

1. *A high-quality computer generated voice is comparable to the voice of a professional narrator concerning (a) learning outcome, and (b) acceptance*

2. *Picture driven e-learning modules with voice-over on small screen mobile devices have the same learning outcome as their standard screen equivalent on desktop or laptop computers*

The learning module Cystic Fibrosis was refactored to meet the experimental requirements to be tested with 107 high school students: (1) the learning content complexity was adapted for K12 (high school) biology lessons; (2) the learning content length was shortened from 111 to 41 pages fitting a 20-minute learning phase; (3) the navigation concept was simplified: all non-linear navigation elements (i.e. content-tree and picture overview) were deactivated and blanked out so that only "next page", "previous page" and "read again" remained available.

Refactoring toward K12-compatibility took place in the mentioned TTS-supporting (Text-To-Speech) CMS which works with artificial voices of different TTS-vendors. In a pilot survey 6 voices were rated by 95 students of the University of Bern concerning understandability and likeability. The winner – a female voice from a TTS-Vendor from Germany – has been implemented. According to Linked et al. [Linked, 09] there is no speaker/gender effect in multimedia messages. Nevertheless, the human voice equivalent (artificial voice-over) was female, too.

2 Method

2.1 Participants

107 subjects (84 females), ranging in age from 16 to 23 years (M = 18.05; SD = 1.09) participated in this experiment. All participants were German speaking K12 students at the Gymnasium Solothurn (Switzerland) where the experiment was conducted.

After the experiment, participants were presented with chocolate. They were tested to be unbiased with regard to the hypotheses under investigation.

2.2 Design

A nested design with the following grouping variables was conducted: (1) Information presentation mode (human voice-over vs. artificial voice-over vs. written text (on screen) instead of spoken information) and (2) screen size (small vs. standard screen size). Participants were parallelized by gender and randomly allotted to four experimental groups shown in table 1.

	Information presentation mode		
Group	***Written text***	***Voice-over***	***Screen-size***
A (n = 26)	*yes*	*none*	*standard (15")*
B (n = 26)	*none*	*human*	*standard (15")*
C (n = 26)	*none*	*artificial*	*standard (15")*
D (n = 29)	*none*	*human*	*small (2.8")*

Table 1: Content disposition modalities

2.3 Apparatus

The e-learning module was either running on a Lenovo ThinkPad X40 .The Students accessed the learning content using Internet Explorer in full-screen mode (groups A to C) or on a HTC Hermes 100 smart-phone using mobile versions of the mentioned browser (full-screen mode, group D). Participants operating the ThinkPad navigated by clicking the links "next page" or "previous page" with an external standard mouse, the mobile users used a stick to operate the touchscreen. Participants allotted to the voice-over conditions (groups B to D) additionally used headphones and were offered the navigational element "read again".

2.4 Material

The e-learning sequence was on cystic fibrosis. Cystic fibrosis is a common disease, which affects the entire body, causing progressive disability and often early death.

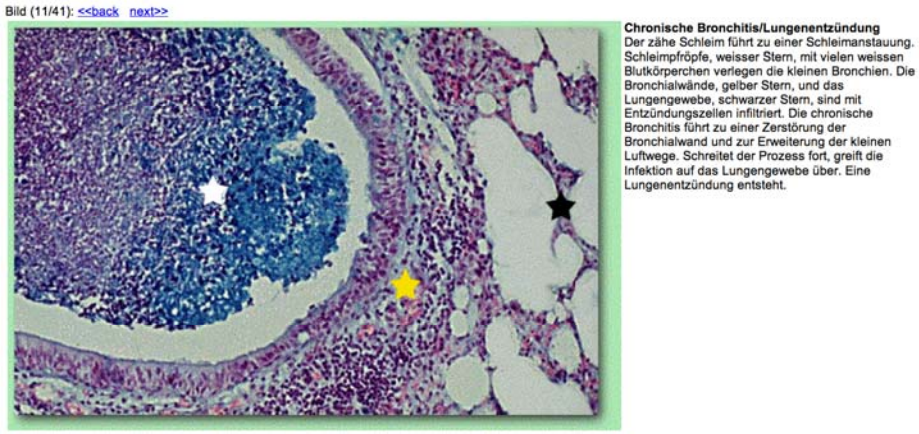

Figure 1: Cystic Fibrosis e-Learning Module: Page #11 standard screen, no voice, written text on the right side of the screen. Translation of the text: Chronic bronchitis/pneumonia. The viscous mucus causes mucosal congestion. Mucosal clots (white star) with many white blood cells displace the bronchioles. The bronchial walls (yellow star) and lung tissue (black star) are infiltrated by inflammatory cells. Chronic bronchitis leads to the destruction of the bronchial wall and to expansion of the small air passages. If the process progresses, the infection will affect the lung tissue as well, causing pneumonia.

2.5 Measurements

2.5.1 Control variables

2.5.1.1 Perceived self-efficacy toward ICT

Compeau and Higgins [Compeau, 95a] state that when individuals have higher self-efficacy toward ICT (Information and Communication Technologies), they also feel that ICT is more useful. Additionally, Compeau and Higgins [Compeau, 95b] and Compeau et al. [Compeau, 99] indicated that when subjects have higher self-efficacy toward ICT, they intend to use ICT more often. The variable was measured by the following item: "How do you rate your ability to use computers and the Internet?". The item was answered on a 4-point scale ranging from *1 = I usually have problems using the computer/Internet* to *4 = I never have problems using the computer/Internet*.

2.5.1.2 Prior knowledge

Six yes/no-questions were asked in order to detect how big the subject's prior knowledge of the learning subject was (e.g., "In the past, I have already informed myself on cystic fibrosis"). So, every participant could reach a prior knowledge score ranging from 0 to 6 points. Furthermore, after the six-week test break (see task and procedure), all participants were asked if they had informed themselves on the learning subject, i.e. cystic fibrosis, during the break.

2.5.2 Manipulation check

Participants allotted to the voice-over conditions were asked if the voice heard was human or artificial (or "I do not know").

2.5.3 Dependent variables

2.5.3.1 Enjoyment (acceptance)

Enjoyment describes the specification of gratification by using ICT. Enjoyment was measured by one item: "How pleasant was the e-learning module?". The item was answered on a 3-point scale ranging from *1 = unpleasant* to *3 = pleasant*.

2.5.3.2 Likeability of the voice heard (acceptance)

Participants allotted to the voice-over conditions (groups B to D) were asked how likeable the voice heard was like (from *1 = not likeable at all* to *3 = very likeable*).

2.5.3.3 Self-perceived learning outcome

In order to detect to which degree the participants think that they learned during the learning phase, one item was presented: "How much did you learn from the e-learning module?". The answers were collected by using a 3-point scale ranging from *1 = little* to *3 = much*.

2.5.3.4 Obtained learning outcome

A knowledge test on cystic fibrosis was constructed. The test consisted of 20 multiple-choice questions. Each subject could achieve 20 points maximum. Four of the knowledge test items explicitly referred to pictorial information of the e-learning module by showing the associated image and asking a question about it. Therefore, behind the total score, an image-related sub-score was conducted, too.

2.6 Task and procedure

All participants gave written information about their gender, age, perceived self-efficacy toward ICT and prior knowledge of the learning subject, i.e. cystic fibrosis. They then executed the knowledge-test the first time (pre-test). Afterwards, participants received oral instructions. The experimenter made sure the participants understood the task: attentively watching, reading and / or listening to the presentation. Then, the students completed the 20 minutes learning phase by working through the e-learning module. Afterwards, participants rated their self-perceived learning outcome and then executed the knowledge test as post-test. Additionally, participants within the voice-over conditions rated if the voice heard was (a) human or (b) artificial. Furthermore, these subgroups reported how likeable the voice within the e-learning module was. Finally, six weeks after the session, all participants executed the knowledge test for the last time (late-test). Before, they were asked if they have been engaged in the learning subject by them selves.

3 Results

3.1 Prior knowledge / self-efficacy toward ICT

Overall, subjects reported *no* to *little* prior knowledge of cystic fibrosis (*M* = .58, *SD* = .97) and *rather high* to *high* self-efficacy towards ICT (Information and Communication Technologies) (*M* = 2.73, *SD* = .89). Both, prior knowledge and self-efficacy towards ICT, did not differ between the groups ($F < 1$). No participant had informed himself on the learning subject, i.e. cystic fibrosis, during the break.

3.2 Enjoyment

Participants reported high enjoyment using the e-learning module (*M* = 2.55, *SD* = .61). A one-way ANOVA did not show any differences between groups ($F = 1.37$, $p > .05$).

3.3 Manipulation check

More than half of the participants allotted the voice-over conditions (*n* = 81) did correctly identify if the voice was human or artificial: 62.4% correct answers, 15.25% did not know, 22.35% wrong answers.

3.4 Likeability of the voice heard

Overall, participants judged that the voice heard was moderately likeable (*M* = 2.11; *SD* = .56). The likeability ratings significantly differ between groups ($F = 5.45$; $p = .006$). The post hoc Scheffé test revealed that the difference between group C (standard screen-size and artificial voice-over) and group D (small screen-size and human voice-over) was significant at the alpha level .05. Table 2 shows the mean likeability-ratings for each group.

Q: How likeable did you find the voice heard in the e-learning module? (Scale: 1 = not likeable at all, 2 = medium, 3 = very likeable)			
Group	***N***	***M***	***SD***
B. Human voice-over / standard screen-size	26	2.15	.46
C. Artificial voice-over / standard screen-size	24	1.83	.64
D. Human voice-over / small screen-size	24	2.33	.48

Table 2: Means of the voice-likeability-ratings separated per voice-over condition

3.5 Self-perceived learning outcome

The self-perceived learning outcome was highest in the human voice-over with standard screen-size condition and lowest in the written text with standard screen-size condition (see table 3). A one-way ANOVA did not show significant differences between groups (F = 2.58, p > .05). Moreover, self-perceived learning outcome is positively correlated with the knowledge test score achieved six weeks after the experiment (r = .24, p = .03).

Q: How much did you learn from the e-learning module? (Scale: 1 = little, 2 = medium, 3 = much)			
Group	***N***	***M***	***SD***
A.Written text / standard screen size	25	2.20	.58
B. Human voice-over / standard screen-size	26	2.58	.50
C. Artificial voice-over / standard screen-size	24	2.38	.58
D. Human voice-over / small screen-size	24	2.35	.54

Table 3: Means of self-perceived learning outcome separated per experimental group

3.6 Measured Learning outcome

In order to prove effects on obtained learning outcome, we computed a 3 (*within factor*: pre, post, and late-test) × 4 (*between factor*: groups A to D) ANOVA. Results show that the within factor is highly significant (F = 17.41, p = .00). Table 4 shows the mean scores for each group. No other effects have been found (all F < 1).

Mean scores of the knowledge test (score-range from 0 to 20 pionts)			
Group (23 ≤ *n* ≤ 29)	***Pre-test***	***Post-test***	***Late-test***
A. On-screen text only	7.42	15.21	11.38
B. Human voice-over / standard screen-size	7.24	15.84	12.52
C. Artificial voice-over / standard screen-size	7.12	15.83	12.39
D. Human voice-over / small screen-size	7.03	14.61	11.72

Table 4: Mean scores of the knowledge test over time separated per group

Mean scores of the knowledge tests' pictorial items (score-range from 0 to 4 points)			
Group (23 ≤ *n* ≤ 29)	***Pre-test***	***Post-test***	***Late-test***
A. written text only (on screen)	1.08	3.12	1.67
B. Human voice-over / standard screen-size	1.08	3.62	2.26
C. Artificial voice- over / standard screen-size	1.00	3.52	2.17
D. Human voice-over / small screen-size	.97	2.93	1.69

Table 5: Mean scores of the knowledge test's pictorial items over time for each modality

We computed another 3 × 4 ANOVA in order to check effects on learning outcome regarding the four pictorial items of the knowledge test (items that can only answered with information carried by the pictures). Results show that the produced knowledge scores significantly vary depending on test trial (pre, post and late-test) (F = 193.46, p = 00). Furthermore, the scores significantly differ between groups (F = 3.56, p = .018). Post hoc Scheffé tests revealed that only the difference between group B (standard screen-size and human voice-over) and group D (small screen-size and human voice-over) was significant at the alpha level .05. Table 5 shows the mean scores for each group.

4 Discussion

Both hypotheses, (1) that presenting pictures with both human and artificial voice-over leads to the same learning outcome, and (2) that the e-learning module leads to the same factual learning outcome and acceptance independent of devices and their screen sizes (15" laptop versus 2,8" smart-phone) are generally supported by the findings in this study. There are differences in aspects of pictorial learning (memorizing information only present in an image) and attitude toward the voice heard. While the enjoyment in the e-learning module was well rated with every condition, the artificial voice-over on standard screen was perceived as less likeable than the human voice on the mobile device. The latter result may be interpreted in terms of the importance of the voice-over on smaller screens. Furthermore, participants allotted to the small screen condition scored less regarding the questions comprising pictorial information.

However, the learning outcome with pictures and text (no voice-over) on 15"-screens was not significantly different from the one with pictures and voice-over on the same screen. Did no modality effect-induced split attention occur, or was the effect too low to impact the learning outcome? Schmidt-Weigand [Schmidt-Weigand, 10b] says that the attention splitting is not an interference in the visual-spatial working memory, but an effect of visual attention, in other words, the learner has to decide when to look at the text and when at the image. Hegarty and Just [Hegarty, 93] detected that learners study diagrams mainly after having read full sentences, meaning that picture learning is text-driven and not vice-versa. Under this view, the split attention effect can also be modulated by the Gestalt principles (closeness of the text to the picture) and by didactically appropriate use of language (textual content consistent with pictorial content) [Behrens, 98]. The Gestalt principles cannot be fulfilled when images and text cannot be presented together on one screen, but rather in sequence because of spatial limitations of mobile devices' small screens. In our experimental setup there was - due to the number of test persons available and the obvious drawback of temporal separation of text and picture - no control group experiencing this condition. The experiment at hand did not involve pictorial information, which by its particularly high complexity might increase the probability of split attention induced by the modality effect. The e-learning module contained neither moving images (i.e. animations and films) nor information affecting viewing comfort, such as small details on x-rays, and thus is suitable for displaying on the small screens of mobile devices. However, the increase of factual knowledge for pictorial items was lower in group D (small screen) on a non-significant level (see Table 5).

The question remains how eligible voice-over for the standard screen modality is because it did not foster higher learning outcomes (both with human and artificial voice). Kürschner et al. [Kürschner, 07] discuss two interesting aspects: firstly, that there is a learner's preference for certain modalities, and secondly, that the modality effect on the learning outcome is dependent on the type of learning content. Their findings show that pictorial learning and textual learning (reading text) do interfere when the learning content is of visio-spacial type (e.g. localization of organs in the human body). Mayer [Mayer, 01] does not make such learning content-specific differentiations in his postulation of the modality effect. Furthermore, their test persons stated a preference either for listening voice-over or reading text, and they could choose one of these two modalities but with no effects on the learning outcome. The freedom of modality choice can be seen as an important motivational factor for the learner when he/she is addressed as user of a system: the ISO Standard 9241 Part 110 [ISO, 06] names "Satisfaction" as third main criteria for usability (besides "Effectiveness" and "Efficiency") and in ISO Standard 9241 Part 11 [ISO, 98] "Controllability" is one of 7 requirements, which can be translated as "Freedom of Control". In our experiment the self-perceived learning outcome was highest under the human voice with standard screen-size condition and lowest under the written text on standard screen- size condition, and self-perceived learning outcome was positively correlated with the knowledge test score achieved six weeks after the session. A fourth explanation (besides following Gestalt principles, picture-content/text-content consistence and non-visio-spacial contents) offer Ginns [Ginns, 05] and Tabbers [Tabbers, 02]: in their experiments the modality effect occurred when learning was system-paced but disappeared when learning was self-paced. The e-learning module used in our experiment was self-paced through a simple forward/backward-navigation. Again, we advocate a tight entanglement of system-usability and didactics: this navigation is of high usability and contributes only little to total cognitive load, resulting in a low extraneous cognitive load giving room for germane cognitive load (mental resources allocable to the learning process itself). The cognitive load theory postulates: the less extraneous cognitive load influences the learner, the more mental resources he or she is able to allocate to the learning process [Sweller, 88]. Cognitive load as concept describing limitations of mental resources becomes obvious in extreme situations, e.g. in aviation, split attention through modality effects is a significant factor decreasing cognitive capacity under circumstances of high cognitive load [Orasanu, 02], e.g. when two airplanes are approaching each other too closely. Systems designed to support pilots in these situations (Traffic Alert and Collision Avoidance System (TCAS)) are equipped with voice-over [Ford, 93]. However, multimedia learning is typically not that mission-critical, but when multimedia content is presented in final medical exams (computer based assessment), the cognitive load is turning high in both extraneous and intrinsic aspects.

Our study has been designed to investigate the didactic appropriateness of the management decision of the Medical Faculty of the University of Bern to switch from professional human narrators to computer generated voices in selected e-learning modules for economic reasons. It could be shown that computer generated voices are a time- and cost-saving substitute for human voices and that the picture with voice-over (and no text) combination perfectly fits small-screen mobile devices. However, our study revealed no modality effect-induced split attention resulting in lower learning outcomes in the picture with text (and no voice-over) modality on standard screens, putting the original motivation for voice-overs into question. In our follow-up experiments we will focus on the threshold, where the modality effect-induced split attention impacts the learning outcome by forcing temporal information separation and building increasing spatial information into pictorial learning content. The results shall help to optimize the allocation of e-learning production resources.

Acknowledgements

A big ‘thank you’ goes to Institute of Medical Education staff members Hans Holzherr for the translation into English, and to Moritz Strickler for his support in text processing. We also thank Markus Tanner of the Gymnasium Solothurn and his biology classes for being part of the experiment.

References

[Behrens, 98] Behrens, R. R.: Art, Design and Gestalt Theory: Leonardo, 31(4), 299-303, 1998.

[Compeau, 95a] Compeau, D. R., Higgins, C. A.: Application of social cognitive theory to training for computer skills. Information Systems Research, 6(2), 118–143,1995a.

[Compeau, 95b] Compeau, D. R., Higgins, C. A.: Computer self-efficacy: development of a measure initial test. MIS Quarterly, 19(2), 189–211, 1995b.

[Compeau, 99] Compeau, D. R., Higgins, C. A., Huff, S.: Social cognitive theory and individual reactions to computing technology: a longitudinal study. MIS Quarterly, 23(2), 145–158, 1999.

[Cooper, 98] Cooper, G.: Research into Cognitive Load Theory and Instructional Design at UNSW. Available at http://education.arts.unsw.edu.au/CLT_NET_Aug_97.HTML (last retrieved 27 March 2006), 1998.

[Cromley, 10] Cromley, J. G., Snyder-Hogan, L. E., Luciw-Dubas, U. A.: Cognitive activities in complex science text and diagrams. Contemporary Educational Psychology, 35(1), 59-74, 2010.

[Florax, 10] Florax, M., Plötzner, R.: The influence of presentation format and subject complexity on learning from illustrated texts in biology. In K. Gomez, L. Lyons, J. Radinsky (Eds.), Proceedings of the Ninth International Conference of the Learning Sciences (Vol. I, pp. 17-24). Chicago, IL: International Society of the Learning Sciences, 2010.

[Ford, 93] Ford, T.: Traffic Alert and Collision Avoidance System. Aircraft Engineering and Aerospace Technology, 62(4), 21-23, 1993.

[Ginns, 05] Ginns, P.: Meta-Analysis of the modality effect. Learning and Instruction, 15, 313-331, 2005.

[Hegarty, 93] Hegarty, M., Just, M.A.: Constructing mental models of machines from text and diagrams. Journal of Memory and Language, 32, 717-742, 1993.

[Kürschner, 07] Kürschner, C., Schnotz, W., Eid, M.: Welchen Einfluss haben die Präsentationsmodalität und Modalitätspräferenzen auf die kognitive Verarbeitung von Text mit

Bildern. Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 39(2), 70-83, 2007.

[Kürschner, 08] Kürschner, C., Schnotz, W.: Das Verhältnis gesprochener und geschriebener Sprache bei der Konstruktion mentaler Repräsentationen. Psychologische Rundschau, 59(3), 139-149, 2008.

[ISO, 98] International Organization for Standardization: Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability. International Standard ISO 9241-110:2006(E), 1998.

[ISO, 06] International Organization for Standardization: Ergonomic requirements for office work with visual display terminals (VDTs) - Part 110: Dialogue Principles. International Standard ISO 9241-11:1998(E), 2006.

[Linked, 09] Linked, B., Geujets, P., Scheiter, K.: The speaker/gender effect: does the speaker's gender matter when presenting auditory text in multimedia messages?. Instructional Sciences, 38, 503–521, 2009.

[Mayer, 98] Mayer, R., Moreno, R.: A split attention effect in multimedia learning: Evidence for dual processing systems in working memory. Journal of Educational Psychology, 90(2), 312-320, 1998.

[Mayer, 01] Mayer, R.: Multimedia Learning. Cambridge University Press, Cambridge, United Kingdom, 2001.

[Nguyen, 06] Nguyen, D.P., Guggisberg, M., Burkhart, H.: Multimedia Information and Mobile-Learning. 8th IEEE International Symposium on Multimedia, San Diego, CA., 2006.

[Orasanu, 02] Orasanu, J., Fischer, U., Davison, J.: Risk perception: A critical element of aviation safety. 15th Triennial World Congress, Barcelona, Spain, 2002.

[Rughooputh, 09] Rughooputh, S.D., Santally, M.I.: Integrating text-to-speech software into pedagogically sound teaching and learning scenarios: Education Tech Research 57, 131–145, 2009.

[Schmidt-Weigand, 10a] Schmidt-Weigand, F., Kohnert, A., Glowalla, U.: Explaining the Modality and Contiguity Effects: New Insights From Investigating Students' Viewing Behaviour. Applied Cognitive Psychology, 24(2), 226-237, 2010a.

[Schmidt-Weigand, 10b] Schmidt-Weigand, F., Kohnert, A., Glowalla, U.: A closer look at slit visual attention in system- and self-paced instruction in multimedia learning. Learning and Instruction, 20, 100-110, 2010b.

[Sweller, 10] Sweller, J.: Cognitive Load Theory: New Conceptualizations, Specifications, and Integrated Research Perspectives. Educational Psychology Review, 22(2), 115-121, 2010.

[Sweller, 88] Sweller, J.: Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257-285, 1988.

[Tabbers, 02] Tabbers, H. J.: The modality of text in multimedia instructions: Refining the design guidelines. Unpublished doctoral dissertation. Heerlen, The Netherlands: Open University of the Neetherlands, 2002.

[Yuanchun, 03] Yuanchun S., Weikai X., Guangyou X., Runting S., Enyi C., Yanhua M., Fang L.: The Smart Classroom: Merging Technologies for Seamless Tele-Education. Advancing research and practice in mobile and ubiquitous computing, 03, 1536-1268, 2003.