Fall 11-21-2017

# Effects of Pedagogical Agent Design on Training Evaluation Measures: A Meta-Analysis

Timothy J. Quesnell
*DePaul University*, tquesnel@depaul.edu

# Effects of Pedagogical Agent Design on Training Evaluation Measures: A Meta-Analysis

A Dissertation

Presented in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

By

Timothy Joseph Quesnell

November 20, 2017

Department of Psychology

College of Science and Health

DePaul University

Chicago, Illinois

**Dissertation Committee**

Jane Halpert, Ph.D., Chairperson

Douglas Cellar, Ph.D.

Goran Kuljanin, Ph.D.

Erich Dierdorff, Ph.D.

Peter Hastings, Ph.D.

**Acknowledgements**

The first people I need to thank for the opportunity to pursue a Ph.D. are my family. Without the sacrifices they made, the appreciation for education they instilled in me, and the confidence they gave me to do whatever I put my mind to I would never have made it this far. They trusted the path I was on and supported me no matter what. They spent a large portion of their lives making sure I've had opportunities in mine, for which I couldn't be more grateful.

I also need to thank the friends (read: family) I met at DePaul over the years. The amount of support, cooperation, and camaraderie that was generated within a group of people (who are too smart for their own good) has been unbelievable. The sense of community we created across multiple cohorts with people from all over the world is something you couldn't recreate if you tried. We did grad school right.

And to all my other friends and family members who ever asked me, "How's school going?" and, "Are you done yet?", thank you too. I know your questions always came from a good place, and that even though it took me forever to finish, you were always cheering for me.

Last but not least, I need to thank Dr. Annette Towler, Dr. Jane Halpert, and the rest of my committee for their help, support, and concern during this long journey. Jane took me under her wing when I was looking for guidance, and I knew I could count on the Committee for quick, thoughtful, and meaningful feedback to help make my dissertation project everything it turned out to be. I couldn't have done it without all of you. Thank you all for being on my team.

## Biography

The author was born in Milwaukee, Wisconsin, November 19, 1985. He graduated from Marquette University High School in Milwaukee in 2004. He received his Bachelor of Science degree in Psychology from the University of Wisconsin- Madison in 2008, and a Master of Arts degree in Industrial Organizational Psychology from DePaul University in 2013.

## Table of Contents

## List of Tables

## List of Figures

**Abstract**

Pedagogical agents are, "conversational virtual characters employed in electronic learning environments to serve various instructional functions" (Veletsianos & Miller, 2008). They can take a variety of forms, and have been designed to serve various instructional roles, such as mentors, experts, motivators, and others. Given the increased availability and sophistication of technology in recent decades, these agents have become increasingly common as facilitators to training in educational settings, private institutions, and the military.

Software to aid in the creation of pedagogical agents is widely available. Additionally, software use and agent creation often requires little formal training, affording nearly anyone the opportunity to create content and digital trainers to deliver it. While the popularity of these instructional agents has increased rapidly in practice, it has outpaced research into best practices for agent design and instructional methods.

The personas programmed into pedagogical agents are recognizable by the people interacting with them, and have been shown to impact various learning outcomes. The form and realism of training agents have also been shown to have substantial impacts on people's perceptions and relationships with these beings. Additionally, agents can be designed in environments that utilize different methods of content delivery (e.g., spoken words versus text), resulting in varying levels of cognitive load (and thus, varying learning outcomes). In an educational setting, agent perceptions and interactions could impact the effectiveness of a training program.

This meta-analysis uses the Integrated Model of Training Evaluation and Effectiveness (IMTEE) as an over-arching framework to examine the effects of training characteristics on training evaluation measures (Alvarez, Salas, & Garofano, 2004). Training characteristics refer to any training-specific qualities that may impact learning outcomes compared to other training programs that offer the same or similar content. Training evaluation refers to the practice of measuring important training outcomes to determine whether or not a training initiative meets its stated objectives. The pedagogical agent training characteristics evaluated in this study include agent iconicity (level of detail and realism), agent roles, and agent instructional modalities. The evaluation measures being examined include post-training self-efficacy, cognitive learning, training performance, and transfer performance.

The Uncanny Valley Theory (Mori, 1970) suggests that agent iconicity (level of detail and realism) is expected to relate to training evaluation measures differently for human-like and non-human-like agents, such that low levels of iconicity (high realism) in non-human-like agents and moderate levels of iconicity in human-like agents would result in optimal training outcomes. These hypotheses were partially supported in that trainees achieved the highest levels of performance on transfer tasks when working with moderately realistic human-like trainers. No significant effects were seen for non-human-like trainers. Additionally, it was expected that the relationship between instructional modality and all training evaluation measures would be positive and stronger for modalities that produce deeper cognitive processing (Explaining and Questioning) than the

modalities that produce shallower processing (Executing and Showing). This hypothesis was not supported.

The relationship between agent role and all training evaluation measures was expected to be positive and stronger for modalities that produce deeper cognitive processing (Coaching and Testing) than the roles that produce shallower processing (Supplanting and Demonstrating). This hypothesis was not supported. Additionally, agents that minimize extraneous cognitive processing were also expected to outperform those that require excess cognitive demands. Agents that utilize speech, personalized messages, facial expressions, and gestures were expected to lead to improved training outcomes compared to those that primarily utilize text, speak in monologue, are expressionless, and/or are devoid of gestures. This hypothesis was partially supported in that agents who were merely present on-screen (physically directing learner attention) resulted in the lowest transfer task performance compared to more active agents who delivered actual content (via speech or text). Learner control (versus trainer control) over support delivery was expected to contribute to improved training outcomes, and support that is delayed in its delivery was expected to hinder performance on training evaluation measures. These hypotheses were not supported.

This meta-analysis, backed by an integration of theories from computer science and multiple disciplines within psychology, contributes to the field of employee training by informing decisions regarding when and how pedagogical agents can best be used in applied setting as viable training tools.

# Effects of Pedagogical Agent Design on Training Evaluation Measures: A Meta-Analysis

## Introduction

Pedagogical agents have been defined as "conversational virtual characters employed in electronic learning environments to serve various instructional functions" (Veletsianos & Miller, 2008). The use of conversational virtual characters dates back to 1966, and as technology has improved, the level of sophistication and accessibility of digital trainers has increased (Salas & Cannon-Bowers, 2001; Weizenbaum, 1966). No longer limited to isolated computer science laboratories, software to create pedagogical agents is now available to almost anyone, including educators, the military, and companies seeking to implement technology-driven instruction (TDI) programs.

There are multiple reasons to study the use of pedagogical agents in training. The first reason is the cost associated with instructor-led training scenarios. Most U.S. companies have training programs in place that use human trainers to teach employees the knowledge and skills necessary to be successful on the job. The American Society for Training & Development's (ASTD) 2013 State of the Industry Report estimates that U.S. companies spent over 164 billion dollars on employee learning in 2012 (ASTD, 2013). It has been estimated that, after wages, benefits, implementation costs, materials, and redistribution of human capital, it costs a company an average of $955 to train just one employee (Sugrue & Rivera, 2005). Especially during times of economic downturn, a company may look for ways to improve their bottom line, which often implies

budget cuts and process improvement measures. A wide array of organizational departments and programs could come under evaluation, including employee training programs (Humphreys, Novicevic, Olson, & Ronald, 2010). Given that pedagogical agents have the potential to reduce some of the costs associated with employee training, exploring best practices for their design is essential.

A second reason to study pedagogical agents is to improve the consistency with which training is delivered. There is an array of factors that can impact a traditional person-to-person training program, leading to differences in administration within and between trainers. Examples of these factors include trainer experience, confidence, perceived credibility, and interactions between trainers and learners, or interactions between trainers and the training environment (Swanson & Falkman, 1997). Lack of consistency is a concern because a given training session may leave out important information, or all information may be presented, but in a way that leads to poorer learning outcomes than those elicited via other training methods. Pedagogical agent content delivery is predetermined and programmed, making it well suited to address consistency concerns.

The third major reason to study pedagogical agents is their ability as a training tool to benefit individuals, organizations, and society as a whole (Aguinis & Kraiger, 2009; Arthur, Bennett, Edens, & Bell, 2003). Well-designed training programs (as part of a high-performance work system; HPWS) help build and maintain human capital (e.g., KSAs, motivation, effort, and job performance). In turn, human capital is linked to a variety of positive organizational benefits,

including improved operational performance, profits, growth, and competitive advantage (Becker & Huselid, 1998; Combs, Liu, Hall, & Ketchen, 2006). As individuals and organizations within a society build knowledge and skills, the collective quality of the labor force improves, and with it, the potential for national economic growth (Aguinis & Kraiger, 2009). Clearly, understanding and improving organizational training has significant and far-reaching effects. Considering the potential pedagogical agents have as a training tool, it is important to determine the characteristics and conditions that result in optimal outcomes when they are utilized.

**Training Evaluation and Training Effectiveness**

Over the past few decades, multiple methods have been developed by which training programs can be evaluated. "Training evaluation" is a term often used to describe the practice of measuring important training outcomes to determine whether or not a training initiative meets its stated objectives (Alvarez, Salas, & Garofano, 2004). Whereas training evaluation is a practical (often quantitative) approach to studying training, theoretical frameworks have also been developed for thinking about and describing the factors that impact training results. These frameworks offer explanations of how and when high-level, macro categories of variables impact the outcomes of training, and are often grouped under the term, "training effectiveness". Despite being separate constructs, the two are related in that training effectiveness factors are studied by measuring training evaluation variables (Alvarez et al., 2004).

The Integrated Model of Training Evaluation and Effectiveness (IMTEE), developed by Alvarez et al. (2004), seeks to combine the two constructs (training evaluation and training effectiveness) into one comprehensive model. The IMTEE was developed following a thorough training evaluation and effectiveness literature search and review. The authors then examined relationships between evaluation and effectiveness measures, and created the model presented in Figure 1.

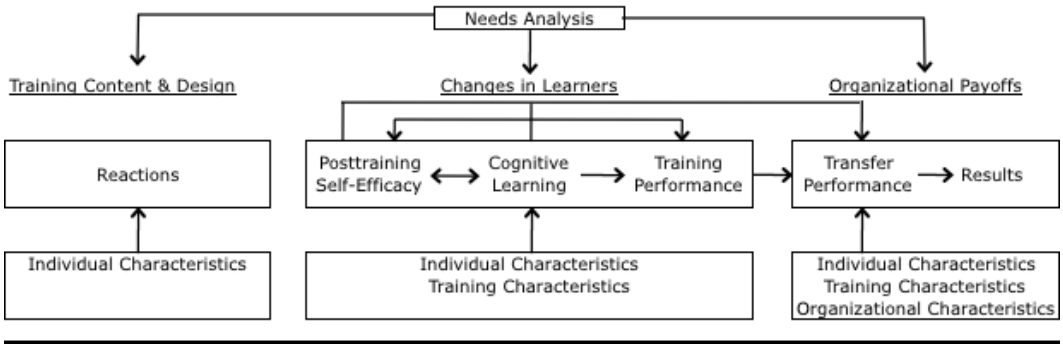


Figure 1: *Alvarez, Salas, & Garofano's (2004) Integrated Model of Training Evaluation and Effectiveness*

Structurally (as can be seen in the model), the IMTEE has four levels, the first of which is Needs Analysis. Needs Analysis is widely accepted as a best practice for developing training content and its design, defining the desired changes in learners, and identifying the eventual organizational payoffs from training (Goldstein & Ford, 2002). Training content & design, changes in learners, and organizational payoffs together make up the second layer of the model. This layer represents the broad categories under which evaluation measures and effectiveness concepts are grouped. The third level of the IMTEE

outlines the evaluation measures identified in the literature as most relevant to evaluating the elements of a Needs Analysis (the second level headings). For example, training content & design can be studied via reactions to the training, while changes in learners can be evaluated by examining post-training self-efficacy (the trainee's belief in his/her ability to perform a specific task after receiving training), cognitive learning (measured immediately after the training to gauge recognition and/or recall of the material presented), and training performance (performance of a relevant task immediately after the training). Finally, potential organizational payoff can be estimated by measuring transfer performance (performance of a novel task at some point after training, where knowledge from the training is required for success) and results. Together, the learning outcome measures included in the third layer of the model are referred to by the authors as the six "targets of evaluation". This model and its six targets is not posited to be exhaustive, but is presented as the most comprehensive and relevant model given the current state of the training evaluation literature (Alvarez et al., 2004). While the third level of the IMTEE specifies how evaluation measures fit into the overall model, the fourth level highlights how the most current and popular effectiveness variable categories (individual, training, and organizational characteristics) are related to training quality, and at which stage of training (before, during, after) these factors can have an impact.

Individual-level training effectiveness factors refer to any learner-specific traits or qualities that may impact learning outcomes compared to other individuals who experience the same training session. An example of individual-

level impact would be learners with high pre-training self-efficacy scoring higher on learning outcomes than learners with low pre-training self-efficacy. Training-level training effectiveness factors refer to any training-specific qualities that may impact learning outcomes compared to other training programs that offer the same or similar content. An example of training-level impact would be a training program that allows users to pause or rewind training videos resulting in improved learning outcomes over a training program that does not allow for the use of pause or rewind features. In the proposed study, differences between pedagogical agents are training-level effectiveness factors that we believe will impact training outcomes. Finally, organizational-level training effectiveness factors refer to qualities or features of the setting in which the training occurs that may impact learning outcomes compared to other settings that offer the same training program. An example of organizational-level impact would be a company that allows employees the time and resources to practice skills presented in a training program achieving better results compared to another company that does not allow practice after the same training (Alvarez et al., 2004).

After analyzing the relationships between training evaluation and training effectiveness measures, Alvarez et al. (2004) found that environmental & organizational characteristics (e.g., positive transfer environment) impact transfer performance and results measures, while training characteristics (e.g., behavioral modeling, practice, feedback on results, etc.) impact transfer performance, results, and all three measures of changes in learners. Additionally, individual trainee

characteristics (e.g., cognitive ability, pre-training self-efficacy, motivation, etc.) can impact all six targets of evaluation (Alvarez et al., 2004).

Burke and Hutchins (2008) provided support for the Alvarez et al. (2004) IMTEE model in their study of best practices for training transfer. They cite the contribution of the Alvarez et al. (2004) model to the understanding of training transfer given the model's emphasis on the "primary transfer influences" of learner characteristics, training intervention design and delivery, and the work environment in which training occurs. In addition to validating the role of these primary factors in training transfer, Burke and Hutchins (2008) wanted to highlight the roles various stakeholders (i.e., trainees, trainers and supervisors) and time periods (i.e., before, during, and after training) can play in training transfer effectiveness.

In their study, Burke and Hutchins (2008) sought to improve the often anecdotal, unfounded, and/or outdated recommendations for achieving or enhancing training transfer in a way that is simultaneously practical and theoretically sound. To this end, the authors gathered 195 unique, written responses (from 92 training professionals) to the following prompt: "Please type a brief statement about what practices you consider effective for supporting training transfer". Thirty-six percent of their participants identified their job title as "training associate" and 30% identified as "managers", with 48% of all respondents in possession of a Master's degree and an average of 14.5 total years of training experience, validating them as subject matter experts (SMEs).

The authors used a quantitative content analysis procedure that allowed them to explore predetermined variables that impact training transfer while allowing emergent themes to be identified. Their predetermined variables of "learner characteristics", "intervention design and intervention delivery", and "work environment" map directly onto the IMTEE effectiveness variables of "individual characteristics", "training characteristics", and "organizational characteristics".

Learner characteristics (individual characteristics) are operationalized as, "attributes regarding the trainee's ability, motivation, personality, perceptions, expectations, or attitudes that influence transfer". Intervention design and intervention delivery (training characteristics) are operationalized as, "the instructor's plan or blueprint for the learning intervention, typically based on needs assessment information and firm goals, or the activities occurring during training delivery". Work environment (organizational characteristics) are operationalized as, "any influence(s) on transfer existing or occurring outside the learning intervention itself [including the evaluation of training transfer]" (Burke & Hutchins, 2008).

Though there is clear overlap between the primary variables Burke and Hutchins (2008) coded and the effectiveness variables of the IMTEE, the authors also coded a category of variables that describe training transfer activities that can occur "before", "during", and "after" training to increase likelihood of training transfer (time periods). While not explicitly called out as a level in the IMTEE model, the IMTEE does implicitly acknowledge the temporal relationships that

exist leading to training transfer (e.g., pre-training self-efficacy impacts cognitive learning, which impacts training performance, which impacts transfer performance). Additionally, Burke and Hutchins (2008) coded the roles of trainees, trainers, and supervisors, which also exist as components of the IMTEE (individual characteristics, training characteristics, and organizational characteristics, respectively) (Alvarez et al., 2004).

Burke and Hutchins (2008) stated that the results of their training transfer best practices study support the Alvarez et al. (2004) training effectiveness categories discussed above. Specifically, experienced training professionals cite the theoretical, primary influences of training transfer (as outlined in the IMTEE) to be critical components of training transfer in practice. Additionally, the stakeholder and time period variables (implicit in the IMTEE and explicit in the elaborated model by Burke and Hutchins (2008)) also revealed themselves to be important factors in training transfer, as identified by professional trainers. In sum, what the Burke and Hutchins (2008) study illustrates is that the Alvarez et al. (2004) IMTEE model (as a synthesis of decades of training evaluation and training effectiveness research) serves as a useful framework for thinking about and modeling training inputs, processes, and outputs (e.g., training transfer and ultimately organizational results), as identified by independent, knowledgeable, and experienced SMEs.

The IMTEE was chosen as the framework for the current study for multiple reasons. One reason is that the evaluation criterion level of the model is a synthesis of multiple influential evaluation models presented throughout recent

decades. These synthesized models include Kirkpatrick's four-dimensional measurement typology (i.e., reactions, learning, behavior, and results) (Kirkpatrick, 1976), the expansion of the Kirkpatrick typology (adding post-training attitudes and training & transfer performance as divisions of behavior) by Tannenbaum, Cannon-Bowers, Salas, and Mathieu (1993), Holton's three evaluation targets of learning, transfer, and results (Holton, 1996), and the multidimensional target areas for training evaluation (content/design, changes in learners, and organizational payoffs) described by Kraiger (2002). A second reason for choosing this model is it acknowledges that a comprehensive review of training programs includes the six targets of evaluation discussed within the effectiveness criteria described above. The final reason for using this model is that the IMTEE is founded on both theory and sound psychometrics (Alvarez et al., 2004; Burke & Hutchins, 2008). While this meta-analysis will not examine the individual-level or full environmental-level effectiveness criteria of the IMTEE, the study will serve as a starting point for a more comprehensive review using the full model.

The basic premise of the current study is that a wide array of design options for pedagogical agents exist (and have been implemented), and the options a programmer chooses during implementation may impact important results and outcomes of the training. Given this, the learner-level evaluation measures outlined in the IMTEE (post-training self-efficacy, cognitive learning, and training performance) and the organizational-level evaluation measure of transfer performance will serve as the criteria by which differences in pedagogical

agents (training effectiveness factors) will be analyzed. The ways by which agents can vary are discussed in more detail below.

**Flexibility in Agent Design**

Pedagogical agents have been programmed to serve a variety of roles, from instructor, to learning partner, to mentor. Not only have these roles been intentionally programmed, but the people interacting with these agents can perceive their various roles and ascribe different attributes to them. Each role can be defined with its own nuances and subtleties, which learners can differentiate. For example, agents have been programmed to serve as "mentors", "experts", and "motivators", each with unique influences on learning and learner motivation (Baylor & Kim, 2004). Pedagogical agents have even been programmed to exude charisma, a trait commonly reserved for the most likeable and adept speakers, lecturers, and "social butterflies" (Towler, Arman, Quesnell, & Hofmann, 2014). The flexibility and range of pedagogical agent designs make them both interesting, and particularly vulnerable to suboptimal design. The study seeks to examine the following elements that can be programmed into pedagogical agents, potentially impacting their effectiveness: 1) degree of human likeness, 2) degree of agent iconicity (level of detail and realism), and 3) pedagogical agent instructional style.

**Agent Aesthetics, Human Likeness, and Iconicity**

Aesthetics are an important component of pedagogical agent design. Similar to human-human interactions, people quickly develop first impressions and stereotypes based on the outward appearance of pedagogical agents. These

initial impressions can subsequently impact learning outcomes, such as information recall (Veletsianos, 2010). The outward appearance of a pedagogical agent can impact perceptions of the agent's role, the characteristics attributed to the agent, and can guide the types of interactions learners have with the agents (Baylor & Kim, 2005). One popular convention for the design of pedagogical agents is to make them increasingly humanlike. The rationale behind this trend is that, if these agents may be used to replace human-delivered training, then they should look as human as possible in appearance, movement, and emotion. Additionally, the technology to design humanlike agents is readily available, removing a major barrier to their creation.

Bates (1994) argues that we should strive for "believability" any time we create a digital character. The believability of a character is the level to which an agent "provides the illusion of life". This illusion of life is the fundamental element that allows people to connect with and be influenced by a non-living character. He posits that, only when it appears agents have desires and interests do people attend to those priorities and make them their own. In a training context, this means trainees would ideally adopt the same values as the trainer, increasing the amount of intrinsic motivation devoted to learning the content being trained, thus improving outcomes of the training (Colquitt, LePine, & Noe, 2000). One would rationalize then that the most "believable" type of character to deliver a training program in an organization potentially seeking to replace or avoid human trainers would be a humanlike pedagogical agent. Indeed, many researchers and practitioners have followed this line of reasoning, creating very realistic digital

trainers. However, there is some evidence that improper agent design can negatively impact learning outcomes, especially when using very humanlike agents.

Mori (1970) first described the notion of the "Uncanny Valley", a popular concept in robotics and medical prosthetic aesthetics fields. The Uncanny Valley (visualized in Figure 2) is the theory that people react more positively (as measured by comfort or familiarity) to non-human agents as they become more humanlike.

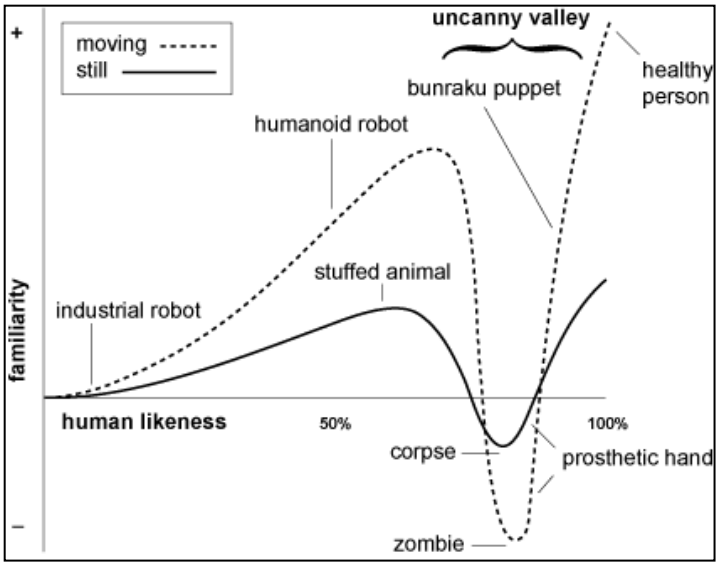


Figure 2*: Mori's (1970) Hypothesized "Uncanny Valley"*

This relationship between human likeness and agent ratings is proposed to be positive, until the agent's design reaches a point whereby it becomes too real, and subjective opinions of the agent decline quickly and significantly. Additionally, if the agent in question is programmed to move, the curve of the Uncanny Valley is magnified (which is especially relevant for pedagogical

agents). Examples of negative reactions cited by Mori (1970) include the surprise elicited by unexpectedly shaking a prosthetic hand believed to be real, or the eeriness of zombies, who appear to be alive and quite human, but fall just short into the Uncanny Valley. Mori's theory then states that as an agent surpasses the Valley, evolving to become an actual human figure, peoples' reactions improve sharply.

There are different expectations and assumptions associated with agents who fall on different points of the Uncanny Valley curve. At the low-human likeness end of the curve exist items like industrial robots, perhaps those that work on assembly lines or in foundries. The expectations associated with these robots are that they are programmed, lifeless, predictable, and perform a limited set of predetermined functions. Moving up the curve, one can imagine encountering more humanlike robots, those with more distinct human features like eyes or hands, or legs that allow them to walk. Some may be programmed with voices and personalities as well, which can be perceived and differentiated by those who interact with them. However, agents in the mid-range of the Uncanny Valley curve possess and display robotic or fictional characteristics, making it apparent that they are not actually alive, and limited by nature of being a robot. The combination of familiar, humanlike features and obvious programmed, robotic limitations creates a realistic set of user expectations. Users assume the agent has a certain amount of advanced ability associated with the visible human characteristics, but the clear robotic components prompt the users

to temper their expectations, and to approach the agent realistically with regard to its abilities.

The high-human-likeness end of the Uncanny Valley curve features agents that are incredibly human-like and realistic. These agents may interact with the environment around them, have human voices, exhibit smooth, realistic movements, and be presented in high definition, or be made of natural-looking synthetic hair and skin. The initial high fidelity of these agents may elicit high expectations for the users, leading them to assume the agent is capable of information processing and social interactions that they are not actually capable of executing. When these assumptions are challenged, possible reactions include repulsion, rejection, confusion, and at the very least, distraction from the task at hand. Though people tend to treat technology in a social manner, interacting with agents so close to the edge between human and non-human could create a form of cognitive dissonance. More specifically, when one's beliefs about the interaction do not align with what is actually happening, the result could be uneasiness and discontent. Reeves and Nass (1996) have even suggested that the human brain hasn’t evolved to process this balance between technology and real social interaction, which would make learning from agents that exist in this middle-ground more difficult than learning from agents whose characteristics better align with our expectations.

One caveat to this discussion is that not all pedagogical agents fall on a continuum book-ended by the categories of “robot” and “human”. Some agents, for example, are designed as paperclips, bugs, or animals. This study contends

that all non-human agents exist toward the lower ("robotic") end of the Uncanny Valley curve. The rationale is that the anthropomorphic qualities associated with training delivery juxtaposed to the electronic, mechanical, programmed qualities associated with computer-generated training programs creates the same realistic training expectations, whether the agent is a robot or a digital insect.

Pedagogical agents (no matter their form) tend to exist in a narrow band of the Uncanny Valley. They are not typically designed at the lowest end of the curve, to be industrial and lifeless in appearance (which would make it difficult to deliver any type of instruction), nor do they exist at the high end of the curve, in the physical environment (as humans and humanoid robots exist). Gulz and Haake (2006) have described a useful typology for categorizing the appearance of pedagogical agents. They argue that agents can vary with regard to their degree of iconicity, or the "degree to which a depicting representation is simplified and reduced" (Gulz & Haake, 2006). Figure 3 illustrates the examples of iconicity cited in their article.

Figure 3*: Examples of Iconicity (Gulz & Haake, 2006)*

Gulz and Haake (2006) provided the images in Figure 3 as examples of their dimension of iconicity-realism. On the left is the most iconic image, and on the right is the most realistic. As pedagogical agents, we would expect these figures to fall along different points on the Uncanny Valley curve, with the leftmost agent being the least human-like, and the rightmost agent being the most human-like. As such, we would also anticipate differential reactions to and expectations of each figure. These differential reactions and expectations could impact the agent's effectiveness at delivering training content to learners.

In his book examining the design and impact of comic book characters, McCloud (1993) argues that when people interact with other social beings, they tend to look directly at the other actor, and therefore, have a very detailed mental representation of that actor. The representation of the social other is realistic. People also maintain a mental representation of themselves during social interactions, however, the image of themselves is much more iconic. Therefore, McCloud (1993) believes that as agents become more iconic, they more easily generate identification and social affinity, thus increasing their impact on users (in the case of training, this impact is learning).

Taken together, there appears to be some confusion regarding best practices for incorporating realism and human-likeness in pedagogical agents. On one hand, pedagogical agents designed to be too robotic, lifeless, or non-human may fall short of generating the social cues necessary to be effective learning aides. On the other hand, designing agents to be too human-like may approach the Uncanny Valley, generating a distraction or negative reactions to the agents, thus

decreasing their effectiveness as trainers. Therefore, one question this research seeks to answer is what level of pedagogical agent iconicity is the "right" level to create the best interaction and most impact. These effects were examined in both human-like pedagogical agents, and non-human-like agents.

A main pitfall of the Uncanny Valley is that agents become so human-like that they become creepy and distracting. When pedagogical agents are intentionally designed to resemble humans, low levels of iconicity are believed to hinder the agent's effectiveness. This study posits that human-like agents high on iconicity will not generate the social cues and identification necessary to aid learning, while human-like agents that are low on iconicity will be distracting or "not quite human enough".

> **Hypothesis Ia**: The relationship between human-agent iconicity and performance on all training evaluation measures is an inverted U-shape, such that very low and very high iconicity leads to poorer performance on training evaluation measures.

Additionally, this study posits that the effects of iconicity are different for pedagogical agents intentionally designed to not resemble humans. In these cases, low levels of iconicity may contribute to increased "illusion of life", "believability" of the character, and learner connection to the agent and the material to be learned.

**Hypothesis Ib**: The relationship between non-human-agent iconicity and performance on all training evaluation measures is negative and linear, such that high iconicity leads to poorer performance on training evaluation measures, and low iconicity leads to better performance on training evaluation measures.

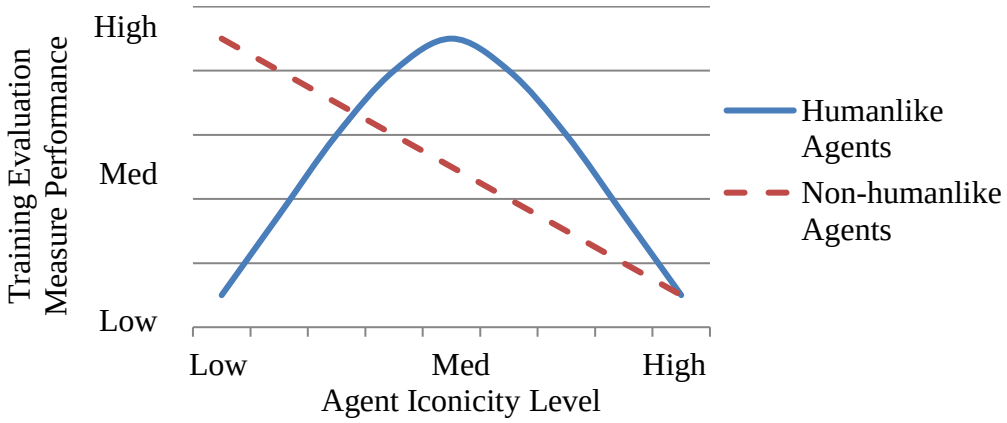


Figure 4*: Hypotheses 1a and 1b: Proposed Relationships Between Iconicity and Ratings.*

**Instructional Design and Social Agency**

Baylor (2000) states that, for pedagogical agents to be effective mentors and trainers, they must display regulated intelligence, exhibit some persona, and display pedagogical control. In her article, she differentiates between "adaptive functionality" and the "agent metaphor". Adaptive functionality is the component of pedagogical agents that allows them to act intelligently, adaptively, and responsively to the learner's actions. The agent metaphor is simply the visible presence of an agent in a learning program (i.e., the portrayal of an animate being). While the importance of the agent metaphor (appearance) has been

addressed in the previous section, the adaptive functionality (or interactive ability component) of pedagogical agents also requires attention.

Many studies have examined pedagogical agents from a technological design perspective (Graesser, Wiemer-Hastings, Wiemer-Hastings, & Kreuz, 1999; Johnson, Rickel, & Lester, 2000; Johnson, Rickel, Stiles, & Munro, 1998; Lester, Voerman, Towns, & Callaway, 1999). Studies of this type focus on the abilities and limitations of the technology behind pedagogical agents. However, attention has recently shifted toward pedagogical agent instructional design (Clarebout, Elen, Johnson, & Shaw, 2002). Studies examining agent instructional design are critical because, just like human instructors, the behaviors, teaching styles, and instructional methodologies of pedagogical agents can have an impact on learning outcomes. However, the work that has been conducted in this area lacks cohesion and common language (Clarebout et al., 2002).

Clarebout et al. (2002) have developed a system for studying, evaluating, and discussing pedagogical agents from an instructional design perspective. Their definition of pedagogical agents is, “animated characters designed to operate in an educational setting for supporting or facilitating learning” (Clarebout et al., 2002; Shaw, Johnson, & Ganeshan, 1999). Given the emphasis on supporting and facilitating learning, the authors refer to their system as a “support typology”. They cite a need for a common language to describe and study pedagogical agents, and created this typology to fill that need. To develop the typology, the authors borrowed from the learning support dimensions described by Elen (1995).

These dimensions include the amount, topical object, formal object, delivery system, and timing of support. The amount of support describes the degree to which learners need assistance during training, and varies according to multiple individual-level characteristics. The topical object of support describes the element of a task being supported (e.g., content or problem solving strategies). The formal object of support describes the elements of the student being supported (e.g., the student's prior knowledge or motivation). The delivery system dimension describes the modality through which learning is supported (e.g., books, teachers, or technological tools), and the timing of support describes at which point the training is delivered (e.g., just-in-time information or delayed feedback).

Clarebout et al. (2002) also describe six different roles agents can play in the delivery of training. These roles include Supplanting (the agent performs most tasks for learners), Scaffolding (the agent performs only the tasks learners cannot perform), Demonstrating (the agent performs a task and then observes the learner perform the task), Modeling (the agent demonstrates a task, but articulates the rationale and strategies being used to execute the task), Coaching (the agent provides hints and feedback when the learner has trouble executing a task), and Testing (the agent challenges the learner's knowledge about elements of a task to facilitate learning).

The authors further group qualities and strategies of these roles into "modalities" of support. These modalities include Executing (the agent performs actions instead of the learner performing them), Showing (the agent provides

demonstrations for the learner, later allowing the learner to replicate), Explaining (the agent provides feedback or clarifications about a task while learners perform them), and Questioning (the agent asks questions about the task or elements of the task for learners to answer). The cross table in Table 1 illustrates the relationships between agent instructional roles and modalities.

The overall support typology integrates elements of the instructional roles and modalities mentioned above. The final typology allows for agent categorization using the following characteristics: Instructional Modality (Executing, Showing, Explaining, Questioning), Agent Role (Supplanting, Scaffolding, Demonstrating, Modeling, Coaching, and Testing), Support Object (content, problem-solving, meta-cognition, and technology), Delivery Modality (speech, text, monologue, personalized, facial expressions, gestures), Source of Control (agent or learner), and Timing of Support (prior to the learning task, just-in-time, or delayed). Table 2 provides a useful visual representation of the final dimensions and their descriptions.

Table 1: *Cross Table of Agent Instructional Roles and Modalities (Clarebout, et al., 2002)*

| | | Modalities | | | |
|---|---|---|---|---|---|
| | | Executing | Showing | Explaining | Questioning |
| Roles | Supplanting | ■ | | | |
| | Scaffolding | ■ | ■ | | |
| | Demonstrating | | ■ | | |
| | Modeling | | ■ | ■ | |
| | Coaching | | | ■ | ■ |
| | Testing | | | | ■ |

To highlight the utility value of their support typology, Clarebout et al. (2002) coded multiple examples of pedagogical agents active in the literature (see Table 3). The list of agents they coded included: *Adele* (Ganeshan, Johnson, Shaw, & Wood, 2000; Johnson et al., 2000; Shaw et al., 1999), *AutoTutor* (Graesser et al., 1999), *Cosmo* (Lester, Voerman, et al., 1999), *Gandalf* (Cassell & Thorisson, 1999), *Herman the Bug* (Lester, Stone, & Stelling, 1999), *Jacob* (Evers & Nijholt, 2000), *PPPersona* (Andre, Rist, & Muller, 1999), *Steve* (Johnson et al., 2000), and *WhizLow* (Gregorie, Zettlemoyer, & Lester, 1999; Johnson et al., 2000).

Table 2: *Support Typology (Clarebout, et al., 2002)*

| Support Typology Dimension | Description |
|---|---|
| **Agent Role** | **Level of learning support provided by the agent** |
| Supplanting | Learners observe while agent performs the task (no learner action) |
| Scaffolding | The agent performs only the tasks a learner cannot yet perform while learners practice a task |
| Demonstrating | Agent performs example task, allows learner to replicate |
| Modeling | Agent performs a task with explanation of the reasoning process |
| Coaching | Agent provides hints/feedback while learner is performing the task |
| Testing | Agent asks learner questions about the task to guide learning |
| **Instructional Modality** | **Methods of conveying successful task completion** |
| Executing | Task is performed by the agent for the learner (no learner action) |
| Showing | Executing, but learner performs task after |
| Explaining | Agent provides task clarification while learner performs the task |
| Questioning | Agent asks questions about the task for learner to answer |
| **Support Object** | **Components of the task agent is targeting to support** |
| Content | Specific elements of the subject matter/topic |
| Problem-Solving | Strategies used to solve a problem or complete a task |
| Meta-cognition | Highlighting learning goals, monitoring learning progress, and evaluating learning strategies |
| Technology | Support related to technology or tools used to complete a task |
| **Delivery Modality** | **Method of communication from agent to learner** |
| Speech | One form of verbal communication |
| Text | A second form of verbal communication |

| | |
|---|---|
| Monologue | Agent talks to the learner, but does not engage in dialogue |
| Personalized | Dialogue between learner and agent |
| Facial Expressions | One form of non-verbal communication |
| Gestures | A second form of non-verbal communication |
| **Control** | **Specifies whether trainer or trainee initiates agent support** |
| Agent | One possible initiator |
| Learner | The other possible initiator |
| **Support Timing** | **The point during which the agent provides support** |
| Prior | Before the learner attempts to solve a task |
| Just-In-Time | As a learner attempts to solve a task |
| Delayed | After the learner has attempted a task |

Table 3: *Analysis of Different Pedagogical Agents (Clarebout, Elen, Johnson, & Shaw, 2002)*

| Analysis of the different pedagogical agents | **Modality** | | | | **Role** | | | | | **Object** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Executing* | *Showing* | *Explaining* | *Questioning* | *Supplanting* | *Demonstrating* | *Modeling* | *Coaching* | *Testing* | *Content* | *Problem solving* | *Metacog-nition* | *Techno-logy* |
| ADELE | | | X | X | | | | X | | X | X | | X |
| STEVE | | X | X | | | X | X | X | | X | X | | X |
| HERMAN | | | X | | | | | X | | X | X | | |
| COSMO | | | X | | | | | X | | X | X | | |
| WHIZLOW | X | | X | | X | | | X | | X | | | |
| PPPERSONA | | X | | | | | X | | | X | | | |
| JACOB | | X | X | | | | X | X | | | X | | |
| GANDALF | | X | X | | | | X | X | | X | | | |
| AUTOTUTOR | | | X | X | | | | X | X | X | | | |
| ***Total*** | ***1*** | ***4*** | ***8*** | ***2*** | ***1*** | ***4*** | ***1*** | ***7*** | ***1*** | ***8*** | ***5*** | ***0*** | ***2*** |

| | **Adaptation** | | **Delivery modality** | | | | | | **Control** | | **Timing** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Quantity* | *Object* | *Speech* | *Text* | *Monologue* | *Personalized* | *Facial expressions* | *Gestures* | *Agent* | *Learner* | *Prior* | *Just-in-time* | *Delayed* |
| ADELE | X | X | X | X | X | X | X | X | X | X | | X | X |
| STEVE | X | X | X | | X | X | X | X | X | | X | X | |
| HERMAN | X | X | X | | X | X | X | X | X | | | X | |
| COSMO | X | X | X | | X | X | X | X | X | X | | X | |
| WHIZLOW | X | | X | | X | X | | X | X | X | | X | |
| PPPERSONA | | | X | | X | | X | X | X | | | | |
| JACOB | X | | | X | | X | | X | X | X | | X | |
| GANDALF | X | | X | | X | X | X | X | | X | | X | |
| AUTOTUTOR | | | | X | X | X | X | X | X | | | | |
| ***Total*** | ***7*** | ***4*** | ***7*** | ***3*** | ***8*** | ***8*** | ***7*** | ***9*** | ***8*** | ***5*** | ***1*** | ***7*** | ***1*** |

Though similarities exist between the agents coded in Table 3, it becomes clear that different pedagogical agents have been designed to exhibit differences on the support typology dimensions (even amongst the nine agents coded). For example, WhizLow is the only agent to exhibit an instructional modality of Executing, while only Steve serves the role of Demonstrating. All agents except for Jacob focus on Content as their object of support and only half of the agents exhibit Quantity and Object adaptations. The delivery modality seems to be the most consistent dimension across agents, but variation does exist. The same is true for support timing and control.

With a tool for describing pedagogical agents in hand, it is important to consider how these different attributes could impact the desired outcomes of training. Differences in instructional design as defined by the support typology may elicit differences in levels (or depth) of processing.

**Levels of Processing**

The Levels-of-Processing theory is a learning theory first put forth by Craik and Lockhart (1972) in an effort to explain how learning and memory are achieved through cognitive encoding. They argue that different types of encoding (mental processes that act on information) range in depth from “shallow” to “deep”. A critical component of this theory is that the deeper information is processed, the more likely it is to be “encoded” into a stronger, more elaborate, and more persistent “memory trace” (which other researchers might refer to as “Long-Term Memory”) (Broadbent, 1958; Craik & Lockhart, 1972).

This theory gained support from a series of studies conducted by Craik and Tulving (1975). The basic premise of the studies was to present words to participants and ask them to interpret the words using varying levels of processing. The varying levels of processing were elicited using the following types of questions about the words (from shallow to deep): 1) an analysis of physical structure of the word (e.g., does the word have 5 letters?), 2) a phonemic analysis of the word (e.g., does the word rhyme with "step"?), or 3) a semantic analysis of the word (e.g., is the word a type of automobile?). Semantic analysis was also induced using sentence completion tasks (e.g., Does the word fit into the following sentence: "The boy walked to the_____"). The participants were then asked to recognize and/or recall as many words as possible (Craik & Tulving, 1975).

Results of the studies provided strong support for the Levels-of-Processing theory. First, it appears that it takes individuals longer to process more abstract questions about the words (which the authors interpret as increased elaboration of the information, and increased cognitive activity). Second, recognition of words increased significantly from words evaluated for physical structure (shallow processing) to words evaluated for phonemic characteristics. Additionally, recognition of words increased significantly from words evaluated for phonemic characteristics to words evaluated for semantic characteristics (deep processing). Craik and Tulving (1975) thus concluded that words paired with deeper processing resulted in better memory traces for those words than those words processed with more shallow tactics.

The experiments conducted by Craik and Tulving (1975) also illustrated that these processing effects on memory also occur with free recall memory tasks and after either expected or unexpected memory tests. In addition, they showed that these effects are stronger when the target words make logical sense within the context of the questions (i.e., the statements are "congruous", thus allowing learners to create unified, elaborated, and deeper mental connections) (Schulman, 1974). Finally, in a study separating response latency (i.e., processing time) from the actual depth of processing, the authors found that the act of processing information at a deeper level appears to be the cause of these effects, not necessarily the amount of time spent doing so (Craik & Tulving, 1975).

Taken together, the results from these Levels-of-Processing studies are important to the current study. Various pedagogical agent interaction styles, as defined by the support typology, could logically result in varying levels of cognitive processing, and thus, varying levels of memory and content learning. For example, of the four instructional modalities in the Clarebout et al. (2002) support typology (Executing, Showing, Explaining, Questioning), pedagogical agents programmed to use Executing and Showing tactics for extended periods of time require participants to passively absorb information as it is presented. Though Showing may require participants to demonstrate what was presented after the training, there is no action or additional information processing requested of the learners "in the moment" or during the training, when encoding of information is likely to occur.

Explaining and Questioning, however, ask participants to take a more active role during learning, which would require deeper processing of the information to be learned. Under an Explaining modality, for example, learners receive instruction and clarification as they struggle to apply the information they receive. Similarly, when agents utilize a Questioning modality, they ask participants to think critically about and make connections with information that has been presented, answering questions about the material throughout the training session. Asking participants to process information contemporaneously as they learn it, whether through applied problems or responding to relevant questions, will likely lead to deeper cognitive processing, and subsequently, enhanced learning.

> **Hypothesis II**: The relationship between instructional modality and all training evaluation measures is positive and stronger for modalities that produce deeper cognitive processing (Explaining and Questioning) than the modalities that produce shallower processing (Executing and Showing).

Similarly, the roles pedagogical agents can assume are likely to encourage varying levels of processing on the part of learners. Coaching and Testing roles tend to rely heavily on the Explaining and Questioning modalities discussed above. Coaching involves explanations and clarifications as learners are actively applying new information to problems, and a Testing role utilizes the practice of

Questioning. Alternatively, Supplanting and Scaffolding require much less effort and activity on the part of learners. Information is transmitted to learners via passive observation, with no further encouragement to process or deeply encode the information at the time of presentation.

According to the support typology, another possible set of roles an agent can assume is that of Modeling or Demonstrating, during which the agent utilizes the passive methods of Supplanting and Demonstrating to display and demonstrate information for learners. However, they also do a better job of explaining the rationale and thought processes involved than these more passive roles. The added insight and clarity defined by the Modeling and Demonstrating roles may require learners to make an increased number of connections between new material and their current knowledge. Additionally, hearing new information and seeing it performed and explained by a Model or Demonstrator may result in increased information elaboration and encoding versus hearing it alone, however, to a lesser degree than other, more active methods of learning. Therefore, it is expected that agents programmed to have Coaching and Testing roles will produce better learning outcomes in trainees than agents in Supplanting and Scaffolding roles. Further, Modeling and Demonstrating agents should elicit a level of processing and elaboration higher than the Supplanting/Scaffolding agents, but lower than the Coaching/Testing agents.

**Hypothesis III**: The relationship between agent role and all training evaluation measures is positive and stronger for those that produce deeper cognitive processing (Coaching and Testing) than those that produce shallower processing (Supplanting and Scaffolding). As roles that produce a moderate level of processing, Demonstrating and Modeling will fall between the other four groups with regard to learning outcomes.

## Cognitive Overload and Multimedia Learning

The cognitive demands of information processing have long been a concern for Industrial/Organizational psychologists. For example, Feldman (1981) discusses the importance of cognitive processes with regard to performance appraisals. He argues that the process of categorizing (or mentally grouping) stimuli is a basic tenet of perception, information storage, and organization. When the stimuli in question are workers and their behaviors, categorization can impact employee evaluations. The more easily an individual worker can be assimilated into a supervisor's existing category prototypes for workers, the more likely that the categorization process will be executed automatically (with little to no cognitive resources). Then, the more consistent an employee's behavior is with the supervisor's expectations, the more likely it is that the behaviors will be stored automatically, as corroboration for the category. When the time comes for performance appraisals, if an individual and their behaviors were observed and stored automatically, the appraisal is most likely to be colored by the supervisor's category prototype (as opposed to reflecting the

actual behavior of the employee), resulting in appraisal inaccuracies (Feldman, 1981).

Thus, it may appear that (for the interpretation and recall of employee behaviors) controlled, thoughtful processing (and avoidance of category prototypes) would be the goal for all supervisors. However, while increased attention and thought can lead supervisors to make careful and meaningful connections between employees, their behaviors, and job performance, Feldman (1981) points out that the controlled categorization process is subject to contextual and perceptual factors that lead to categorization errors and eventual evaluation inaccuracies.

This discussion by Feldman (1981) is relevant to pedagogical agent training for a few reasons. The first is that it highlights the difference between automatic and controlled cognitive processes. While some situational characteristics lend well to automatic processing (requiring few cognitive resources), others (as would be the case when attempting to assimilate new knowledge into existing storage) require more effort and cognitive attention. The second reason is that it shows how effortful cognitive processing is imperfect and subject to errors. When these errors manifest, it is typically at a later date during recall (as would be the case during a knowledge test, or on-the-job performance). While Feldman's (1981) work addresses the automatic and controlled nature of cognitive processing, other researchers have dived deeper into controlled processing, and how the quantity of stimuli to be processed can lead to storage errors. The following pages discuss how presentation methods and the quantity of

information presented can overload controlled processing routes, which in turn can lead to suboptimal learning conditions during pedagogical agent training scenarios.

As discussed above, pedagogical agent image, role, and instructional modality can impact learners and learning outcomes. Another important element of the agent-learner training scenario is the communication medium through which interactions occur between the two (Mayer & Moreno, 2003). Traditional educational materials (e.g., textbooks and lectures) and other instructional practices that present messages through one channel only are often based on the information delivery view of learning. This teaching perspective suggests people learn by simply adding new information to what they already know, and that to teach, trainers need only provide information through the verbal channel (Mayer, 2003). Therefore, according to the information delivery view, instruction that occurs solely via written or spoken word should be sufficient. However, this view is inadequate and inconsistent with how people actually learn, as presenting information in such a narrow manner often leads to shallow processing, forgetting of key points, and poorer learning outcomes (Mayer, 2005).

A substantial amount of research has been conducted in cognitive psychology on the dual channel perspective of human information processing (Baddeley, 1992, 1998; Clark & Paivio, 1991; Paivio, 1986). This theory states that there are two channels through which information can be processed: visual and auditory. The visual channel processes information presented in the form of images or animations, while the verbal channel processes either spoken words or

printed text. A growing body of research suggests processing that occurs simultaneously through both channels is likely to lead to deeper processing and improved learning outcomes than processing through one channel alone. This dual processing is referred to as “multimedia learning” and is especially relevant to pedagogical agent training (Mayer, 2001).

Especially important with regard to pedagogical agents is another concept known as the “modality effect”. This refers to the idea that learning can be improved if information presented in text is instead presented in an auditory format with visual support, such as graphs, diagrams, or animations. The modality effect may support and can help improve the effectiveness of multimedia learning (Ginns, 2005).

The Cognitive Theory of Multimedia Learning (Mayer, 2001) relies heavily on the premises of multimedia learning discussed above, and specifies three assumptions about information processing that are relevant to this pedagogical agent discussion. The first assumption is the dual channel assumption, outlined above. Again, one channel (the eyes/visual channel) receives and processes visual stimuli/information, while a second channel (the ears/auditory channel) receives and processes verbal stimuli/information. The second assumption is the limited capacity assumption, which states that the ability to process information in either channel is limited. This limit implies that when demands on cognitive resources in either channel are too great, a person may be forced to pay attention to certain information while neglecting other information (Mayer, 2005). The third assumption, the active processing assumption, states that

deep learning occurs when the learner is able to pay attention to and select important information being presented, organize the information into meaningful visual and auditory representations, and combine them with what is already known. The end result of active processing is the ability to problem solve, utilizing the newly acquired information.

Given these assumptions, the Cognitive Theory of Multimedia Learning posits that learners can engage in three types of processing: 1) essential processing, 2) incidental processing, and 3) representational holding. Essential processing is cognitive processing required for selecting, organizing, and integrating the material to be learned. Incidental processing is cognitive resources being devoted to extraneous information presented in addition to the required materials. Finally, representational holding is cognitive resources being devoted to holding mental representations in working memory over a period of time. Therefore, when learners attempt to learn information, the total cognitive processing power required is the sum of essential processing, incidental processing, and representational holding. “Cognitive overload” occurs when the amount of processing required is more than the amount of cognitive resources the learner possesses.

Mayer and Moreno (2003) cite five types of avoidable cognitive overload that can occur, which reduces the amount of deep processing and learning experienced by the learner. The most applicable type of overload for the current study occurs when one channel is overloaded with essential processing demands. For example, on-screen text appears concurrently with the animation the text is

describing. This creates what is known as the split-attention effect (Ginns, 2005; Towler et al., 2008). The learner has to split his/her attention between what he/she is seeing and what he/she is reading, which causes him/her to only be able to select some information to process through the working memory (Mayer & Moreno, 2003). As social beings, pedagogical agents elicit attention from learners, requiring learners to expend cognitive resources in doing so. Therefore, when a training scenario requires learners to pay attention visually to the agent or materials (e.g., charts, graphs, diagrams) while simultaneously read and process text, a large amount of information is likely to flood one (visual) information processing channel. Such a scenario would make it unlikely that the learner could effectively process the content to be learned, leading to shallow processing and stunted learning outcomes.

> **Hypothesis IVa**: The relationship between delivery modality and all training evaluation measures is positive and stronger for agents that utilize speech as the primary delivery mechanism than those that use text as the primary mechanism.

An important delivery modality distinction made in the support typology by Clarebout et al. (2002) is between monologues and personalized messages. By the authors' definition, training sessions can be described as monologues when agents are the sole communicators, receiving no input from the learners, providing little opportunity for social exchange, and talking to (rather than with)

learners. Agents who use a personalized delivery modality, on the other hand, establish a dialogue with learners, providing information, receiving feedback (in one form or another), and reacting to this feedback with an appropriate response. While a monologue delivery modality more closely aligns with outdated learning theories (i.e., with learners as passive recipients of information), personalized delivery modalities engage learners and encourage participation, which may lead to deeper information processing. Additionally, interaction and feedback exchanges between agents and learners contribute to the social nature of a training scenario. Again, if learners perceive the training agent to be more lifelike and believable, it may lead to improved learning outcomes over more one-sided training programs.

> **Hypothesis IVb**: The relationship between delivery modality and all training evaluation measures is positive and stronger for agents that utilize personalized messages than those that use monologues.

Along similar lines, agents who effectively and naturally use facial expressions and gestures to convey ideas and information are more likely to be perceived as lifelike and believable beings in a social learning context. Thus, agents who exhibit these qualities should elicit improved learning outcomes over agents who do not.

**Hypothesis IVc**: The relationship between delivery modality and all training evaluation measures is positive and stronger for agents that utilize facial expressions to help deliver their message than those that do not utilize facial expressions.

**Hypothesis IVd**: The relationship between delivery modality and all training evaluation measures is positive and stronger for agents that utilize gestures to help deliver their message than those that do not utilize gestures.

Another question addressed in the support typology is the initiation of support during the training session. Agents can be programmed to offer supplemental assistance at various points throughout the training session, reacting to various user actions with a support response. In situations such as these, the agent is said to have control over the support delivery. Alternatively, some training programs offer users the option to essentially pause the training to ask for help when it is needed. In these situations, the learner is said to have control over support delivery. Providing this freedom and control over the pace of the training is likely to engage learners, and the amount of elaboration provided by the support is likely to lead to deeper processing on the part of the learners. For these reasons, learner control over support delivery is likely to lead to better outcomes than agent-controlled support delivery.

**Hypothesis V**: The relationship between support control and all training evaluation measures is positive and stronger for training scenarios that allow learner control over support delivery than those that require the agent to determine when support is delivered.

Finally, the support typology acknowledges that the timing of support delivery is an important factor in training scenarios. Providing support before a training session is likely to have two important impacts. The first is that it may prime learners for information to come, such that they are more attuned to the information when it is presented. The second is that it may lead to increased meta-cognitive activity as learners plan, execute, and monitor their learning activities throughout the training. Support that is provided during the training (just-in-time) allows learners to exhaust their personal cognitive resources in an attempt to resolve learning tasks on their own. Allowing users to work through problems on their own with minimal assistance is likely to produce deeper cognitive processing and improved learning outcomes. Additionally, just-in-time support is more likely to lead to immediate application of the information (and therefore allows less time to forget it).

**Hypothesis VI**: The relationship between support timing and all training evaluation measures is negative and stronger for training scenarios that primarily provide support after learners attempt a new task (delayed) than those that primarily provide support prior to or during (just-in-time) learners attempting a new task.

By definition, pedagogical agents exist to offer support during training sessions. As noted above, the support typology suggests that there are four main targets of their support: content, problem-solving, meta-cognition, and technology. It is an assertion of this study that, in cases where support directed toward any of these areas is useful, an offering of any of these types of support would benefit the learners in the form of improved learning outcomes. However, the IMTEE provides no recommendations as to which type of support might result in the most benefit to learners. Given that training designers are likely to use training supports judiciously (favoring the training content to emphasizing support mechanisms, or at least striking a balance between the two), there is merit to exploring which types of training supports lead to better learning outcomes than others. In this study, this exploration will come in the form of Research Question 1.

**Research Question 1**: Does focusing on any of the four objects of support (i.e., content, problem-solving, meta-cognition, technology) result in improved learning outcomes more so than focusing on other objects of support?

Table 4 provides a concise summary of the hypotheses and research questions addressed by this study.

Table 4: *Summary of Proposed Hypotheses and Research Questions*

| Hypothesis / Research Question | Description |
|---|---|
| Hypothesis Ia | The relationship between *human*-agent iconicity and performance on all training evaluation measures is an inverted U-shape, such that very low and very high iconicity leads to poorer performance on training evaluation measures. |
| Hypothesis Ib | The relationship between *non-human*-agent iconicity and performance on all training evaluation measures is negative and linear, such that high iconicity leads to poorer performance on training evaluation measures, and low iconicity leads to better performance on training evaluation measures. |
| Hypothesis II | The relationship between instructional modality and all training evaluation measures is positive and stronger for modalities that produce deeper cognitive processing (Explaining and Questioning) than the modalities that produce shallower processing (Executing and Showing). |
| Hypothesis III | The relationship between agent role and all training evaluation measures is positive and stronger for modalities that produce deeper cognitive processing (Coaching and Testing) than the roles that produce shallower processing (Supplanting and Scaffolding). As a role that produces a moderate level of processing, Demonstrating Modeling will fall between the other four groups with regard to learning outcomes. |

| | |
|---|---|
| Hypothesis IVa | The relationship between delivery modality and all training evaluation measures is positive and stronger for agents that utilize speech as the primary delivery mechanism than those that use text as the primary mechanism. |
| Hypothesis IVb | The relationship between delivery modality and all training evaluation measures is positive and stronger for agents that utilize personalized messages than those that use monologues. |
| Hypothesis IVc | The relationship between delivery modality and all training evaluation measures is positive and stronger for agents that utilize facial expressions to help deliver their message than those that do not utilize facial expressions. |
| Hypothesis IVd | The relationship between delivery modality and all training evaluation measures is positive and stronger for agents that utilize gestures to help deliver their message than those that do not utilize gestures. |
| Hypothesis V | The relationship between support control and all training evaluation measures is positive and stronger for training scenarios that allow learner control over support delivery than those that require the agent to determine when support is delivered. |
| Hypothesis VI | The relationship between support timing and all training evaluation measures is negative and stronger for training scenarios that primarily provide support after learners attempt a new task (delayed) than those that primarily provide support prior to or during (just-in-time) learners attempting a new task. |
| Research Question 1 | Does focusing on any of the four objects of support (i.e., content, problem-solving, meta-cognition, technology) result in improved learning outcomes more so than focusing on other objects of support? |

A meta-analysis in this domain to address these questions is warranted for a few reasons. The body of literature related to pedagogical agents currently feels

scattered and incohesive; there is a substantial amount of primary research being conducted regarding the appearance and behaviors of pedagogical agents, but no clear overarching framework to describe pedagogical agents, think about their design, and link their various attributes to their outcomes. Generally, most pedagogical agent labs seem to be following their own agenda as opposed to contributing to a cohesive body of work. The advancement of technology in recent decades and the proliferation of easy-to-use pedagogical agent software has contributed to the production of pedagogical agents outpacing the research behind them. This is evidenced by the hundreds of different pedagogical agents that all vary from each other in terms of looks and actions.

This meta-analysis attempts to apply a theoretically sound structure to the world of pedagogical agent appearance, instructional behaviors, and social behaviors, and to see how well the results coincide with that structure. Specifically, the study seeks to synthesize the array of forms (e.g., human-like/non-human-like, realistic/iconic) pedagogical agents have taken on, to validate the use of the Clarebout et al. (2002) support typology to categorize agent instructional behaviors, and to begin the discussion of what other social, anthropomorphic elements of agent design may be important to consider. The quantitative nature of measuring performance after training makes meta-analysis a more appropriate summary than a narrative review. While this study alone may not result in a single framework within which all pedagogical agents should be considered, it provides a step in the right direction toward unifying the literature relating to pedagogical agent design.

## Method

To explore the hypotheses listed above, meta-analytic techniques were used to look across studies that vary on the dimensions mentioned. The following sections elaborate on the definition of "pedagogical agents" used in this study, operationalize the variables explored, and outline the methodology used to collect and analyze the data.

### Pedagogical Agent Definition

For this study, the functional definition of pedagogical agents was "conversational virtual characters employed in electronic learning environments to serve various instructional functions" (Veletsianos & Miller, 2008). For the purposes of this study, "conversational" referred to an agent's ability to deliver (verbal or written) information to a learner, regardless of whether or not the agent can receive information (e.g., commands, feedback, questions, etc.). Another important element of this definition is the "learning environment". While an agent's environment can take many forms, this study focused on only those agents designed to strengthen the knowledge, skills, or abilities of learners (i.e., serving an "instructional function"). This excluded agents designed purely for entertainment, therapeutic roles, or other non-educational functions. Only studies that presented one agent at a time were included.

### Human Likeness and Iconicity

Given the expected differences between human-like and non-human-like agents, it was important to differentiate between the two. For the purposes of this study, "human-like" agents were defined as those whose form reasonably

approximated that of human beings. This included an evaluation of all physical information available for each agent included in the study. If a full body was displayed to trainees, the agents were coded based on variables such as body proportions, gait, movement, and posture. Additionally, the head and face of each agent was of primary concern as these body parts are presumed to be visible in nearly all training scenarios, and a large portion of human-to-human social cues are displayed and interpreted through the face. Elements of the head/facial region examined were whether all common facial features were included (e.g., two eyes, eyebrows, nose, mouth, ears, etc.), all facial features were within reasonable human proportion and arrangement, and the agent displayed reasonable and appropriate movements, gaze, and eye contact. If any single element of an agent's body did not qualify as distinctly "human-like" (e.g., a Cyclops or a superhero), it was coded as non-humanlike. In addition to physical and non-verbal agent features, it was important to consider the agents' voice (such as speech patterns and intonations) where applicable. For example, if an agent closely resembled a human, but exhibited robotic intonations or unusual speech patterns, learners were continually reminded that the agent was non-human, and was coded appropriately.

Human and non-human pedagogical agents were coded for iconicity based on the prototypes presented in Gulz and Haake (2006). Three levels of iconicity were coded: low, moderate, and high iconicity. Agents coded as having a low level of iconicity ("realistic") were photorealistic, video animated, had high levels of detail in their animation, and/or incorporated high levels of fine lines and shading. Agents coded as high on iconicity were cartoon-like, exhibited unnatural

coloration, movement, lack of detail, blurriness, or other features that made it apparent the agent was created, and was not real. All agents who exhibited equal amounts of these features, or did not fit clearly into one of these categories were coded as “medium iconicity”.

**Article Search**

The article search had multiple components, the first of which was an online database search. The primary researcher (and author of this paper) searched the following databases for relevant articles: Academic Search Complete, Business Source Complete, Computers & Applied Sciences Complete, Education Research Complete, ERIC, Health and Psychosocial Instruments, ProQuest Dissertations & Theses Global Full Text, PsychARTICLES, PsychINFO, Social Sciences Citation Index, and Google Scholar. The search terms used were “Pedagogical ‘and’ Agent*”, “Digital ‘and’ Train*”, “Computer Mediated Training”, “Computer ‘and’ Train*”. The date ranges for the above search engines were from 1960 to March of 2016. This range was chosen to include the initial appearance of pedagogical agents cited above (ELIZA; Weizenbaum, 1966). The sole exception was the Social Sciences Citation Index search, which extended from 1985 through March 2016 (1985 is the earliest date catalogued in this resource).

The second facet of the article search was to comb recently published journals for articles inaccessible on the Internet. A list of relevant journals was composed based on the articles discovered during the online search. The primary

researcher then browsed these journals' article lists from the past twenty years to identify articles that qualified for the study.

Finally, to help address the "file drawer effect", or a bias toward journals publishing articles that achieve significant results, (Rosenthal, 1979), email addresses were obtained for all authors of all studies included in the dataset. The primary researcher also contacted other researchers who have conducted research in this area, but whose articles were not included in the dataset. Messages were sent to these authors to request unpublished studies, and explained the high-level purpose of the study, the types of studies of interest, and how to submit them. Next, the researcher searched reference lists of existing pedagogical agent meta-analyses, any available online conference programs, and websites for faculty that regularly publish pedagogical agent research for potential data sources. Additionally, the researcher leveraged relevant professional organizations/networks for data sources and author contact information. These networks included the American Management Association (AMA), the American Society for Information Science and Technology (ASIS&T), the Association for Computing Machinery (ACM), the Association for Talent Development, the Computing Research Association, the International Society for Performance Improvement (ISPI), the Robotic Industries Association (RIA), the Society for Human Resource Management (SHRM), the Society for Industrial and Organizational Psychology (SIOP), and the United States Distance Learning Association (USDLA).

A single Industrial/Organizational graduate student research assistant was recruited to assist with obtaining relevant author emails from the papers identified in the article search process. The primary researcher collected email addresses from the other sources listed above. The research assistant also helped to compile the citations of the articles included in the final dataset. Her work was double checked and proofread by the primary author. To facilitate record keeping of the article search process, a Google Sheets spreadsheet was used and maintained by the primary researcher. The information tracked included search dates, search terms, databases searched, article type, authors, the articles' publication journal, whether the article was a meta-analysis or not, and each article's place within the results that were displayed (e.g., #127 out of 2,861 results). Tracking this information helped to keep the process organized and documented, and allowed for accurate reporting of search results.

**Article Inclusion**

An initial screen was performed on all articles for quality, empiricism, and sample size adequacy. Articles were included in this study if they (1) used a single pedagogical agent to deliver training, (2) provided enough information to code the agent's appearance as human/non-human and level of iconicity, (3) provided the appropriate amount and level of data to compute required statistics, and (4) measured at least one of the training evaluation criteria of interest from the IMTEE (Post-training Self-Efficacy, Cognitive Learning, Training Performance, and/or Transfer Performance). Studies must have sampled a "normal" adult population (as other samples may not accurately reflect the

general working population). Included articles also provided effect sizes or the ability to compute them from the available data.

Multiple publications from the same dataset were treated as one study. All articles that qualified for inclusion in the study were coded for: human-likeness, level of iconicity, support typology dimensions, and the four IMTEE evaluation targets being examined (Post-training Self-Efficacy, Cognitive Learning, Training Performance, and Transfer Performance). These variables and the coding and analysis procedures used in this study are explained in more detail below.

**Coding Procedure and Analyses**

The primary researcher performed all coding of every article included in the final dataset. A code book (including definitions and examples; see Appendix A) was developed and referenced throughout the coding process.

The primary researcher used a private data entry spreadsheet hosted via Google Sheets. The data entry fields matched those outlined in the code book. Once all relevant data was entered into the shared data entry spreadsheet, it was transferred into Comprehensive Meta-Analysis software (CMA; by Biostat, Inc.) for data analysis. This program was chosen for its flexibility with regard to the inputs it accepts (compatible with 100+ data formats) and its power to easily compute and synthesize effect sizes across these different formats. Common data points entered include correlations, means, standard deviations, and sample size.

For each study selected, the independent variables of interest were entered as “subgroups” into CMA. Once each subgroup’s applicable dependent variable data was entered, CMA automatically computed effect sizes across studies

(Hedge's *g*), allowing us to confirm or refute each hypothesis and the research question. Additional detail regarding the analysis process and results of the analyses are presented in the Results section, below.

# Results

## Article Search Outcome

The keyword searches outlined above resulted in 4,871 articles to review. Of those articles, 101 were identified as potentially relevant based on information contained in the abstracts. Of those articles, 41 were coded for relationships between independent and dependent variables (please see Appendix B for a complete list). As many articles report on a series of studies with multiple samples and/or report on multiple outcome variables, the final dataset resulted in 105 data points (a total of 4,051 respondents) across all articles (an average of 2.46 DVs per article).

Primary reasons for exclusion from the initial subset of 101 articles include: Used multiple agents in the training (k=9), described an agent environment but did not test it (k=9), used a non-adult sample (k=8), reported insufficient data (k=7), did not test a DV of interest (k=6), no relevant IVs were measured (k=5), reported on the same sample as a previous article (k=4), agent did not serve an instructional function (k=4), the paper is theoretical (k=3), the agent was a video of a real human delivering the training (k=3), presence/absence of the agent was not tracked (k=1), agent was a robot (k=1). To retrieve the relevant data from the seven articles in which information was missing/incomplete, the authors were contacted via email. As of the time of reporting, no data from these studies has been received.

Additionally, a total of 138 individual pedagogical agent researchers were identified via the search procedures listed above. These researchers were

contacted for unpublished studies using the most recent and up-to-date contact information available on the internet. Of those contacted, 18 researchers responded. Of those messages, 13 (72.2%) researchers indicated they have no unpublished work in the area, 4 (22.2%) indicated that they conducted no further research in the domain, and 1 (5.6%) indicated he no longer has access to any data that may be relevant to the present study.

**Model Selection: Fixed Effect Versus Random Effects**

The analysis in the present study examines the data utilizing a random effects model as opposed to a fixed effect model. The random effects model is the most appropriate of the two given the sampling methodology. Fixed effect models are appropriate when all studies included in the meta-analysis are intended to estimate the same effect size. That is, all studies are identical to each other in terms of sample selection, methodology, and measurement (only the outcome values differ).

A random effects model is appropriate when studies examine samples from multiple populations within the universe of populations, when multiple methodologies are used, or when studies vary based on the tools used to measure outcomes of interest. Given that the article search methodology identified relevant literature that differed from other studies according to at least one of these criteria, the random effects model is most appropriate. Utilizing a random effects model also allows for broader generalization of the results. Most of the studies collected (57/59 unique samples, 96.9%) utilized a student sample in a lab setting. Given

that a primary goal of this analysis is to explain and predict behavior in the domain of workplace training, a random effects model is again most appropriate.

**Publication Bias**

A potential source of error that can have major effects on the quality of the data in a meta-analysis is publication bias. Publication bias exists in a data set when the research that appears in published literature is in some way systematically different from the universe of completed research studies (published and unpublished). One primary way in which this bias may arise is via the "file drawer effect" (as mentioned above) in which it is assumed that studies are more likely to be accepted for publication if the results are significant. If these articles are published, they are typically easier to access, and the meta-analytic sample will be skewed more heavily toward significant findings (Rosenthal, 1979). As noted in the Methodology section, the present study sought to include published and unpublished studies. Eight of the 41 unique documents (19.5%) originated from unpublished sources, including: unpublished doctoral dissertations (5), conference presentations (1), and other unpublished manuscripts (2). For comparison, Borenstein, Hedges, Higgins, and Rothstein (2009) report that on average, only 8% of manuscripts referenced in meta-analyses tend to be unpublished.

Statistical methods exist for estimating the potential for publication bias, and are based on the following assumptions: (a) Large studies are likely to be published regardless of statistical significance because these involve large commitments of time and resources, (b) Moderately sized studies are at risk for

being lost, but with a moderate sample size even modest effects will be significant, and so only some studies are lost here, (c) Small studies are at the greatest risk for being lost (Borenstein, et al., 2009).

The statistical methods for identifying potential publication bias therefore examine the relationship between sample size and effect size. If unexpected relationships do exist within a given sample, they are attributed to the absence of unpublished studies in the data set. Given that the tests examine potential bias in detecting an individual effect size, publication bias analyses in this study have been run for each dependent variable.

The first method for statistically evaluating the presence or absence of publication bias is to compute a Fail-Safe N. While Rosenthal's (1979) Fail-Safe N calculation is of historical significance for popularizing concern regarding publication bias, it suffers from a few drawbacks that the Orwin (1983) Fail-Safe N method addresses. First, Rosenthal's (1979) method ignores the issue of "substantive significance", instead emphasizing statistical significance. That is, it asks how many hidden studies are required to make an observed effect not statistically significant instead of asking how many hidden studies it would take to make the effect practically unimportant. Second, the formula forces the mean effect size in the hidden studies to be zero, when it could theoretically be negative or positive (but lower than the observed effect). Finally, the Rosenthal (1979) Fail-safe N examines *p*-values across studies, as was common at the time. Today, the common practice is to compute a summary effect, and then compute a *p*-value for this effect (Borenstein, et al., 2009). As such, the Orwin (1983) method (which

accounts for these shortcomings) is what is used in the present study. The purpose of this analysis is to help determine how many relevant studies would need to exist (and not be included already) to reduce the mean effect size to practical insignificance. It answers the question of whether or not the observed effects are entirely due to publication bias instead of the hypothesized relationships.

The results of the Orwin's (1983) Fail-safe N analysis are presented in Table 5. The first row of Table 5 lists the observed Hedge's g (effect size) for each dependent variable. The second row indicates what hypothetical Hedge's *g* value we would consider to be "trivial", or substantially different such that we would draw a different conclusion than the observed Hedge's g. Previous meta-analysis authors have selected "trivial" cut points of 0.10 (Jansen, Daams, Koeter, Veltman, van den Brink, & Groudiaan, 2013; Yildiz, Vieta, Leucht, & Baldessarini, 2011), and some have used cutoffs as relaxed as 0.01 or 0.001 (Bem, Tressoldi, Rabeyron, & Duggan, 2015). Given these precedents, the more conservative 0.10 cutoff has been used in the present study.

The third row of Table 5 illustrates the assumption that the mean Hedge's *g* effect size is 0.00 in whatever studies may be missing from our analyses. This assumption indicates that, on average, these supposed missing studies display no effect on the dependent variables (positive or negative). The results of the Orwin's Fail-Safe N analysis is presented on row 4, indicating the number of studies needed (given the parameters we have set) to reduce the effects of the studies that are included in the analyses (row 5) to the set trivial value.

Table 5: *Orwin's Fail-Safe N Results Summary*

| Dependent Variable: | Post-Training S.E. | Cognitive Learning | Training Performance | Transfer Performance |
|---|---|---|---|---|
| Hedge's g in Observed Studies | .284 | .201 | .250 | .287 |
| Criterion for a 'Trivial' Hedge's g | 0.10 | 0.10 | 0.10 | 0.10 |
| Mean Hedge's g in Missing Studies | 0.00 | 0.00 | 0.00 | 0.00 |
| Number of Missing Studies Needed to Meet "Trivial" Criterion | 32 | 34 | 30 | 66 |
| Number of Studies Collected | 17 | 33 | 20 | 35 |

Note that for each dependent variable analysis, the number of missing studies needed to reduce the effect size to a conservative "trivial" value of 0.10 is more than the number of studies already included in the analysis. This implies that for each learning outcome, for publication bias to make the observed effects "trivial", over half of all relevant studies in existence would have to have been excluded from the analysis. Given the thoroughness of the article search process, this is highly unlikely, and thus we can be fairly certain that publication bias is not the primary driver of the results seen in hypothesis testing.

The next step in this analysis is to estimate what quantity of bias may exist and to estimate what the effect size might be in the absence of this bias. To do so, Duval and Tweedie's Trim and Fill test is employed (Duval & Tweedie, 2000a, 2000b). Trim and Fill is an iterative procedure that removes studies that are outliers (in terms of sample and effect size) one-by-one, at each step re-computing a mean effect size until the remaining studies exhibit a more balanced distribution around the new effect size. The goal of this method is to generate an unbiased estimate of the true effect size. A statistical side effect of this process is that it

yields an artificially narrow confidence interval (since "extreme" values are being removed). To correct for this artifact, the algorithm adds the original studies back into the analysis and imputes a statistical "mirror image" for each to correct the variance (Duval & Tweedie, 2000a, 2000b).

The results of the Trim and Fill analysis are presented in Table 6. An examination of the distribution around the mean effect size for Post-Training Self-Efficacy and Training Performance revealed that no studies would need to be removed to attain an acceptable distribution. However, the removal of 8 studies examining Cognitive Learning and Transfer Performance yielded an improved effect size estimate and distribution around that estimate. This result suggests that the data for these analyses may be skewed in favor of achieving a significant result. Taking together these results, the results of the Fail-safe N tests, and the results of the hypothesis testing (presented below), publication bias was not deemed to be a major influencing factor in the sample of studies collected.

Table 6: *Duval and Tweedie's (2000) Trim and Fill Summary Table*

| | Studies Trimmed | Random Effects Point Estimate | Lower Limit | Upper Limit | *Q*-value |
|---|---|---|---|---|---|
| Post-Training Self-Efficacy | | | | | |
| Observed Values | | 0.271 | 0.116 | 0.426 | 34.6 |
| Adjusted Values | 0 | 0.271 | 0.116 | 0.426 | 34.6 |
| | | | | | |
| Cognitive Learning | | | | | |
| Observed Values | | 0.228 | 0.033 | 0.423 | 193.6 |
| Adjusted Values | 8 | 0.040 | -0.154 | 0.234 | 282.5 |
| | | | | | |
| Training Performance | | | | | |
| Observed Values | | 0.239 | 0.125 | 0.354 | 22.1 |
| Adjusted Values | 0 | 0.239 | 0.125 | 0.354 | 22.1 |
| | | | | | |
| Transfer Performance | | | | | |
| Observed Values | | 0.364 | 0.210 | 0.517 | 125.2 |
| Adjusted Values | 8 | 0.194 | 0.030 | 0.358 | 200.2 |

**Multiplicity, Experiment-Wise Error, & Family-Wise Error**

Similar to individual research studies that conduct multiple analyses on one participant sample, meta-analyses that conduct multiple analyses on a set of research studies are susceptible to an increase in Type I error rates. Unless certain precautions are taken, the more analyses a researcher runs on the data, the more likely he/she is to make a Type I error of rejecting the null hypothesis when it is actually true (a false positive effect). Pigott and Polanin (2014) illustrated that meta-analyses rarely correct for these types of errors despite the fact that this error exists and can have substantial impacts on the conclusions drawn from the data.

Currently, a consensus does not exist regarding the best methods for minimizing the risk of a Type I error. Borenstein, et al. (2009) have argued for corrections as simple as reducing the critical *p*-value from .05 to .01. However, others have argued that this method is purely convention, with no statistical basis (Pigott & Polanin, 2014). Hedges and Olkin (1985), proposed adjusting the alpha level using the equation $\alpha^* = (\alpha/c)$ where $\alpha^*$ is the new critical alpha level, "α" is the original alpha level, and "c" is the number of comparisons being made. This method, when used as an "experiment-wise" correction, however, is susceptible to over-correction as the number of studies included can increase rapidly (making the critical *p*-value quite small). This over-correction results in decreased power to detect a significant result when one actually exists.

Pigott and Polanin (2014) discussed alternatives to this correction, and suggest multiple methods for meta-analysts to minimize the risk of Type I error while simultaneously preserving statistical power. One practice that is relevant to

the current study is to minimize the number of analyses conducted to only those identified *a priori*. As such, the comparisons of interest have been laid out in the Introduction section and will be tested specifically given the data set available.

A second practice the authors advocate that has been utilized in this study is to distinguish between "experiment-wise" and "family-wise" error corrections. While the experiment-wise correction would result in an adjusted critical *p*-value of 0.00048 (.05/105; utilizing every study in the correction), a family-wise correction allows for a more moderate estimate of the critical alpha level by dividing the critical *p*-value by the relevant number of studies involved in each separate dependent variable analysis. For example, if a researcher is analyzing the impact of an independent variable on a dependent variable and 17 of 105 total data points are included in this specific analysis, the new critical *p*-value would be .05/17 = .0029. Similarly, in an analysis of an independent variable/dependent variable relationship in which 8 studies were collected, the corrected alpha level would be (.05/8) = .0063. While the critical *p*-values derived from the family-wise correction are relatively conservative, they are not as overly stringent as the experiment-wise correction would be, better balancing the relationship between Type I error risk and statistical power.

**Hypothesis Testing**

For each hypothesis tested below, the first step was to compute a *Q*-statistic and corresponding *p*-value. The *Q*-statistic is a test of the null hypothesis that variability between studies is due to random error and is not due to real differences between the levels of an independent variable (Borenstein, et al.,

2009). Similar to a significant *F*-value in an ANOVA, a significant *Q*-value indicates that comparisons between the coded levels of the independent variables (referred to as "Moderators" in CMA) may be made. Each *Q*-value that is not significant implies that there are not real differences between the various pedagogical agent characteristics as they relate to the training outcomes of interest. As discussed above, adjusted *p*-critical values have been computed for each effect size and will be used in all hypothesis testing to reduce the risk of committing a Type I error.

Hypothesis Ia stated that the relationship between human-agent iconicity and performance on all training evaluation measures would be an inverted U-shape, such that very low and very high iconicity leads to poorer performance on training evaluation measures. This hypothesis was partially supported, as significant differences exist between levels of iconicity for human-like pedagogical agents when predicting transfer task performance, $Q(2)=18.732$, $p=0.000$. Results are summarized in Table 7.

Table 7: *Hypothesis Ia Overall Results Summary*

| Dependent Variable | Level | *Q*-value | df | *p*-Value | Adj *p* Crit | k |
|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | Overall | 0.082 | 1 | 0.775 | 0.003 | 16 |
| Cognitive Learning | Overall | 9.91 | 2 | 0.007 | 0.002 | 27 |
| Training Performance | Overall | 0.027 | 2 | 0.986 | 0.004 | 14 |
| Transfer Performance | Overall | 18.732 | 2 | 0.000 | 0.003 | 16 |

A significant *Q*-value for transfer performance indicates that significant differences between the levels of iconicity exist for human-like pedagogical agents. As predicted, human-like agents that exhibit moderate levels of realism exhibit enhanced transfer task performance ($Z$=4.182, $p$=0.000) versus agents that are too realistic ($Z$= -1.876, $p$=0.061) or not realistic enough ($Z$=1.698, $p$=0.090). These and all other differences are summarized in Table 8.

Table 8 (and subsequent summary tables) summarize the results for each level of the independent variables' impact on each of the dependent variables. Working from left to right, the tables list the dependent variable of interest ("Dependent Variable"), the level of the independent variable of interest ("Level"), the number of studies included in each comparison ("k"), the observed effect size ("Point Estimate"), the standard error and variance of the observed effect size ("Std Err" and "Variance" respectively), the lower and upper limits of the 95% confidence interval ("Lower Limit" and "Upper Limit", respectively), the results of the *z*-score test of significance ("Z-Score"), and the p-value associated with the *z*-score test ("*p*-value").

Table 8: *Human-Agent Iconicity Results Summary*

| Dependent Variable | Level | k | Point Estimate | Std Err | Variance | Lower Limit | Upper Limit | *Z*-Value | Obs *p*-Value |
|---|---|---|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | High (Cartoon) | 0 | - | - | - | - | - | - | - |
| Post-Training Self-Efficacy | Moderate | 12 | 0.273 | 0.098 | 0.01 | 0.08 | 0.466 | 2.774 | 0.006 |
| Post-Training Self-Efficacy | Low (Realistic) | 4 | 0.326 | 0.157 | 0.025 | 0.019 | 0.633 | 2.079 | 0.038 |
| Cognitive Learning | High (Cartoon) | 6 | 0.736 | 0.240 | 0.058 | 0.266 | 1.206 | 3.067 | 0.002 |
| Cognitive Learning | Moderate | 12 | -0.163 | 0.163 | 0.027 | -0.483 | 0.157 | -1.000 | 0.317 |
| Cognitive Learning | Low (Realistic) | 9 | 0.0248 | 0.181 | 0.033 | -0.108 | 0.603 | 1.366 | 0.172 |
| Training Performance | High (Cartoon) | 5 | 0.243 | 0.151 | 0.023 | -0.052 | 0.538 | 1.614 | 0.107 |
| Training Performance | Moderate | 3 | 0.219 | 0.159 | 0.025 | -0.093 | 0.530 | 1.376 | 0.169 |
| Training Performance | Low (Realistic) | 6 | 0.212 | 0.117 | 0.014 | -0.017 | 0.440 | 1.817 | 0.069 |
| Transfer Performance | High (Cartoon) | 4 | 0.267 | 0.157 | 0.025 | -0.041 | 0.575 | 1.698 | 0.090 |
| Transfer Performance | Moderate | 6 | 0.455 | 0.109 | 0.012 | 0.242 | 0.669 | 4.182 | 0.000 |
| Transfer Performance | Low (Realistic) | 6 | -0.210 | 0.112 | 0.013 | -0.430 | 0.009 | -1.876 | 0.061 |

Hypothesis Ib stated that the relationship between non-human-agent iconicity and performance on all training evaluation measures would negative and linear, such that high iconicity leads to poorer performance on training evaluation measures, and low iconicity leads to better performance on training evaluation measures. This hypothesis was not supported for any of the outcome measures. Results are summarized in Table 9.

Table 9: *Hypothesis Ib Overall Results Summary*

| **Dependent Variable** | **Level** | ***Q*-value** | **df** | ***p*-Value** | **Adj *p* Crit** | **k** |
|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | Overall | - | - | - | - | - |
| Cognitive Learning | Overall | - | - | - | - | - |
| Training Performance | Overall | 0.120 | 1 | 0.729 | 0.008 | 6 |
| Transfer Performance | Overall | 6.865 | 1 | 0.009 | 0.003 | 19 |

Given that none of the observed *Q*-values are significant based on the family-wise adjusted *p*-values, observed differences between pedagogical agents in terms of iconicity for non-human-like agents are likely due to error as opposed to real differences. These differences are summarized in Table 10.

Table 10: *Non-Human-Agent Iconicity Results Summary*

| Dependent Variable | Level | k | Point Estimate | Std Err | Variance | Lower Limit | Upper Limit | *Z*-Value | Obs *p*-Value |
|---|---|---|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | High (Cartoon) | 1 | -0.141 | 0.323 | 0.104 | -0.774 | 0.493 | -0.435 | 0.663 |
| Post-Training Self-Efficacy | Moderate | - | - | - | - | - | - | - | - |
| Post-Training Self-Efficacy | Low (Realistic) | - | - | - | - | - | - | - | - |
| Cognitive Learning | High (Cartoon) | 6 | 0.507 | 0.179 | 0.032 | 0.157 | 0.857 | 2.836 | 0.005 |
| Cognitive Learning | Moderate | - | - | - | - | - | - | - | - |
| Cognitive Learning | Low (Realistic) | - | - | - | - | - | - | - | - |
| Training Performance | High (Cartoon) | 4 | 0.223 | 0.170 | 0.029 | -0.109 | 0.556 | 1.315 | 0.188 |
| Training Performance | Moderate | 2 | 0.317 | 0.209 | 0.044 | -0.094 | 0.727 | 1.513 | 0.130 |
| Training Performance | Low (Realistic) | - | - | - | - | - | - | - | - |
| Transfer Performance | High (Cartoon) | 13 | 0.728 | 0.124 | 0.015 | 0.485 | 0.972 | 5.874 | 0.000 |
| Transfer Performance | Moderate | 6 | 0.174 | 0.171 | 0.029 | -0.162 | 0.510 | 1.016 | 0.310 |
| Transfer Performance | Low (Realistic) | - | - | - | - | - | - | - | - |

Hypothesis II stated that the relationship between instructional modality and all training evaluation measures would be positive and stronger for modalities that produce deeper cognitive processing (Explaining and Questioning) than the modalities that produce shallower processing (Executing and Showing). This hypothesis was not supported for any of the outcome measures. Results are summarized in Table 11.

Table 11: *Hypothesis II Overall Results Summary*

| **Dependent Variable** | **Level** | ***Q*-value** | **df** | ***p*-Value** | **Adj *p* Crit** | **k** |
|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | Overall | 0.678 | 2 | 0.713 | 0.003 | 17 |
| Cognitive Learning | Overall | 3.334 | 4 | 0.503 | 0.002 | 33 |
| Training Performance | Overall | 10.134 | 4 | 0.038 | 0.003 | 20 |
| Transfer Performance | Overall | 10.600 | 4 | 0.031 | 0.001 | 35 |

Given that none of the observed *Q*-values are significant based on the family-wise adjusted *p*-values, observed differences between pedagogical agents in terms of instructional modality are likely due to error as opposed to real differences. These differences are summarized in Table 12.

Table 12: *Instructional Modality Results Summary*

| Dependent Variable | Level | k | Point Estimate | Std Err | Variance | Lower Limit | Upper Limit | *Z*-Value | Obs *p*-Value |
|---|---|---|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | Cannot Determine | - | - | - | - | - | - | - | - |
| Post-Training Self-Efficacy | Executing | 12 | 0.265 | 0.097 | 0.009 | 0.075 | 0.456 | 2.734 | 0.006 |
| Post-Training Self-Efficacy | Showing | - | - | - | - | - | - | - | - |
| Post-Training Self-Efficacy | Explaining | 4 | 0.190 | 0.198 | 0.039 | -0.198 | 0.578 | 0.959 | 0.338 |
| Post-Training Self-Efficacy | Questioning | 1 | 0.471 | 0.282 | 0.079 | -0.081 | 1.023 | 1.674 | 0.094 |
| Cognitive Learning | Cannot Determine | 2 | -0.124 | 0.391 | 0.153 | -0.890 | 0.642 | -0.316 | 0.752 |
| Cognitive Learning | Executing | 21 | 0.198 | 0.127 | 0.016 | -0.051 | 0.447 | 1.556 | 0.120 |
| Cognitive Learning | Showing | 2 | 0.736 | 0.441 | 0.195 | -0.129 | 1.600 | 1.668 | 0.095 |
| Cognitive Learning | Explaining | 6 | 0.173 | 0.242 | 0.059 | -0.303 | 0.648 | 0.712 | 0.477 |
| Cognitive Learning | Questioning | 2 | 0.652 | 0.406 | 0.165 | -0.144 | 1.447 | 1.606 | 0.108 |
| Training Performance | Cannot Determine | 1 | 0.599 | 0.171 | 0.029 | 0.264 | 0.933 | 3.508 | 0.000 |
| Training Performance | Executing | 7 | 0.266 | 0.070 | 0.005 | 0.130 | 0.403 | 3.816 | 0.000 |
| Training Performance | Showing | 3 | 0.395 | 0.164 | 0.027 | 0.074 | 0.716 | 2.410 | 0.016 |
| Training Performance | Explaining | 6 | 0.088 | 0.121 | 0.015 | -0.149 | 0.324 | 0.726 | 0.468 |
| Training Performance | Questioning | 3 | -0.090 | 0.187 | 0.035 | -0.456 | 0.276 | -0.482 | 0.630 |
| Transfer Performance | Cannot Determine | 2 | -0.250 | 0.274 | 0.075 | -0.787 | 0.287 | -0.913 | 0.361 |
| Transfer Performance | Executing | 16 | 0.253 | 0.104 | 0.011 | 0.048 | 0.457 | 2.423 | 0.015 |
| Transfer Performance | Showing | 6 | 0.613 | 0.187 | 0.035 | 0.247 | 0.980 | 3.282 | 0.001 |
| Transfer Performance | Explaining | 7 | 0.472 | 0.166 | 0.028 | 0.146 | 0.789 | 2.840 | 0.005 |
| Transfer Performance | Questioning | 4 | 0.729 | 0.244 | 0.059 | 0.252 | 1.206 | 2.993 | 0.003 |

Hypothesis III stated the relationship between agent role and all training evaluation measures would be positive and stronger for modalities that produce deeper cognitive processing (Coaching and Testing) than the roles that produce shallower processing (Supplanting and Demonstrating). As a role that produces a moderate level of processing, Modeling should fall between the other four groups with regard to learning outcomes. This hypothesis was not supported for any of the outcome measures. Results are summarized in Table 13.

Table 13: *Hypothesis III Overall Results Summary*

| **Dependent Variable** | **Level** | ***Q*-value** | **df** | ***p*-Value** | **Adj *p* Crit** | **k** |
|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | Overall | 3.279 | 3 | 0.351 | 0.003 | 17 |
| Cognitive Learning | Overall | 2.555 | 4 | 0.635 | 0.002 | 33 |
| Training Performance | Overall | 10.576 | 5 | 0.060 | 0.003 | 20 |
| Transfer Performance | Overall | 14.348 | 5 | 0.014 | 0.001 | 35 |

Given that none of the observed *Q*-values are significant based on the family-wise adjusted *p*-values, observed differences between pedagogical agents in terms of agent role are likely due to error as opposed to real differences. These differences are summarized in Table 14.

Table 14: *Agent Role Results Summary*

| Dependent Variable | Level | k | Point Estimate | Std Err | Variance | Lower Limit | Upper Limit | *Z*-Value | Obs *p*-Value |
|---|---|---|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | Cannot Determine | - | - | - | - | - | - | - | - |
| Post-Training Self-Efficacy | Supplanting | 9 | 0.257 | 0.114 | 0.013 | 0.033 | 0.480 | 2.250 | 0.024 |
| Post-Training Self-Efficacy | Scaffolding | 1 | 0.704 | 0.358 | 0.128 | 0.002 | 1.405 | 1.997 | 0.049 |
| Post-Training Self-Efficacy | Demonstrating | 4 | 0.334 | 0.140 | 0.020 | 0.059 | 0.609 | 2.381 | 0.017 |
| Post-Training Self-Efficacy | Modeling | - | - | - | - | - | - | - | - |
| Post-Training Self-Efficacy | Coaching | 3 | -0.020 | 0.230 | 0.053 | -0.470 | 0.431 | -0.085 | 0.932 |
| Post-Training Self-Efficacy | Testing | - | - | - | - | - | - | - | - |
| Cognitive Learning | Cannot Determine | 4 | 0.283 | 0.286 | 0.082 | -0.276 | 0.843 | 0.993 | 0.321 |
| Cognitive Learning | Supplanting | 19 | 0.141 | 0.137 | 0.019 | -0.127 | 0.408 | 1.031 | 0.303 |
| Cognitive Learning | Scaffolding | 5 | 0.467 | 0.274 | 0.075 | -0.071 | 1.004 | 1.703 | 0.089 |
| Cognitive Learning | Demonstrating | - | - | - | - | - | - | - | - |
| Cognitive Learning | Modeling | - | - | - | - | - | - | - | - |
| Cognitive Learning | Coaching | 3 | 0.036 | 0.351 | 0.123 | -0.652 | 0.725 | 0.103 | 0.918 |
| Cognitive Learning | Testing | 2 | 0.652 | 0.414 | 0.171 | -0.158 | 1.463 | 1.577 | 0.115 |
| Training Performance | Cannot Determine | 1 | 0.599 | 0.171 | 0.029 | 0.264 | 0.933 | 3.508 | 0.000 |
| Training Performance | Supplanting | 10 | 0.286 | 0.064 | 0.004 | 0.160 | 0.412 | 4.456 | 0.000 |
| Training Performance | Scaffolding | 1 | 0.334 | 0.286 | 0.082 | -0.227 | 0.894 | 1.168 | 0.243 |
| Training Performance | Demonstrating | 2 | 0.023 | 0.203 | 0.041 | -0.375 | 0.420 | 0.112 | 0.911 |
| Training Performance | Modeling | - | - | - | - | - | - | - | - |
| Training Performance | Coaching | 5 | 0.021 | 0.147 | 0.022 | -0.267 | 0.308 | 0.140 | 0.888 |
| Training Performance | Testing | 1 | -0.150 | 0.264 | 0.070 | -0.668 | 0.367 | -0.570 | 0.569 |
| Transfer Performance | Cannot Determine | 4 | 0.068 | 0.202 | 0.041 | -0.328 | 0.464 | 0.336 | 0.737 |
| Transfer Performance | Supplanting | 14 | 0.220 | 0.118 | 0.014 | -0.011 | 0.451 | 1.866 | 0.062 |
| Transfer Performance | Scaffolding | 5 | 0.987 | 0.221 | 0.049 | 0.555 | 1.419 | 4.476 | 0.000 |
| Transfer Performance | Demonstrating | 5 | 0.259 | 0.186 | 0.035 | -0.107 | 0.624 | 1.388 | 0.165 |
| Transfer Performance | Modeling | - | - | - | - | - | - | - | - |
| Transfer Performance | Coaching | 1 | 0.209 | 0.397 | 0.157 | -0.569 | 0.987 | 0.526 | 0.599 |
| Transfer Performance | Testing | 6 | 0.647 | 0.188 | 0.035 | 0.278 | 1.016 | 3.437 | 0.001 |

Hypothesis IVa stated that the relationship between delivery modality and all training evaluation measures would be positive and stronger for agents that utilize speech as the primary delivery mechanism compared to those that use text as the primary mechanism. During coding of this hypothesis, it was determined that some (2) studies have programmed agents to be present, but did not program them to deliver information via speech or text. Instead, their potential value as agents is derived from their presence, and from gestures and gazes as information is presented to the learners. As such, the coding scheme was adapted to include "present" as an option, in addition to "speech as a primary delivery modality" and "text as a primary delivery modality".

While this hypothesis was not supported as strictly worded (comparing speech to text), a significant effect for transfer performance was detected ($Q(2)=13.165$, $p=0.001$), indicating that significant differences exist between the three levels of delivery modality (effectively speech, text, and "body language") when predicting transfer task performance. Results are summarized in Table 15.

Table 15: *Hypothesis IVa Overall Results Summary*

| Dependent Variable | Level | *Q*-value | df | *p*-Value | Adj *p* Crit | k |
|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | Overall | 0.741 | 1 | 0.389 | 0.003 | 17 |
| Cognitive Learning | Overall | 2.506 | 2 | 0.286 | 0.002 | 33 |
| Training Performance | Overall | 3.628 | 1 | 0.057 | 0.003 | 20 |
| Transfer Performance | Overall | 13.165 | 2 | 0.001 | 0.001 | 35 |

A significant *Q*-value for transfer performance indicates that significant differences exist between the types of delivery mechanisms (text vs. speech vs. body language). As predicted, agents that deliver information primarily via speech exhibit relatively high transfer task performance (*Z*=4.266, *p*=0.000). This result is significant when compared to agents that communicated solely via body language (*Z*=1.426, p=0.154). Similarly, the single study examining the effect of text delivery on transfer performance outperformed the agents that operated without speech or text (Z=4.042, p=0.000). These and all other differences are summarized in Table 16.

Table 16: *Speech versus Text Results Summary*

| **Dependent Variable** | **Level** | **k** | **Point Estimate** | **Std Err** | **Variance** | **Lower Limit** | **Upper Limit** | ***Z*-Value** | **Obs *p*-Value** |
|---|---|---|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | Present | - | - | - | - | - | - | - | - |
| Post-Training Self-Efficacy | Text | 2 | 0.041 | 0.279 | 0.078 | -0.506 | 0.588 | 0.146 | 0.884 |
| Post-Training Self-Efficacy | Speech | 15 | 0.291 | 0.083 | 0.007 | 0.129 | 0.454 | 3.511 | 0.000 |
| Cognitive Learning | Present | 2 | 0.712 | 0.387 | 0.150 | -0.047 | 1.470 | 1.838 | 0.066 |
| Cognitive Learning | Text | 1 | -0.345 | 0.599 | 0.358 | -1.518 | 0.828 | -0.576 | 0.564 |
| Cognitive Learning | Speech | 30 | 0.211 | 0.103 | 0.011 | 0.008 | 0.413 | 2.041 | 0.041 |
| Training Performance | Present | - | - | - | - | - | - | - | - |
| Training Performance | Text | 2 | -0.231 | 0.257 | 0.066 | -0.734 | 0.272 | -0.899 | 0.369 |
| Training Performance | Speech | 18 | 0.269 | 0.054 | 0.003 | 0.163 | 0.375 | 4.973 | 0.000 |
| Transfer Performance | Present | 2 | 0.415 | 0.291 | 0.085 | -0.156 | 0.986 | 1.426 | 0.154 |
| Transfer Performance | Text | 1 | 3.298 | 0.816 | 0.666 | 1.699 | 4.898 | 4.042 | 0.000 |
| Transfer Performance | Speech | 32 | 0.329 | 0.077 | 0.006 | 0.178 | 0.480 | 4.266 | 0.000 |

Hypothesis IVb stated that the relationship between delivery modality and all training evaluation measures would be positive and stronger for agents that utilize personalized messages than those that use monologues. This hypothesis was not supported for any of the outcome measures. Results are summarized in Table 17.

Table 17: *Hypothesis IVb Overall Results Summary*

| Dependent Variable | Level | *Q*-value | df | *p*-Value | Adj *p* Crit | k |
|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | Overall | 0.710 | 2 | 0.701 | 0.003 | 17 |
| Cognitive Learning | Overall | 0.346 | 2 | 0.841 | 0.002 | 33 |
| Training Performance | Overall | 5.335 | 2 | 0.069 | 0.003 | 20 |
| Transfer Performance | Overall | 3.677 | 2 | 0.159 | 0.001 | 35 |

Given that none of the observed *Q*-values are significant based on the family-wise adjusted *p*-values, observed differences between pedagogical agents in terms of agent messaging are likely due to error as opposed to real differences. These differences are summarized in Table 18.

Table 18: *Agent Messaging Results Summary*

| Dependent Variable | Level | k | Point Estimate | Std Err | Variance | Lower Limit | Upper Limit | *Z*-Value | Obs *p*-Value |
|---|---|---|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | Present | 2 | 0.041 | 0.288 | 0.083 | -0.523 | 0.604 | 0.141 | 0.888 |
| Post-Training Self-Efficacy | Monologue | 5 | 0.304 | 0.135 | 0.018 | 0.040 | 0.568 | 2.255 | 0.024 |
| Post-Training Self-Efficacy | Personalized | 10 | 0.282 | 0.114 | 0.013 | 0.058 | 0.506 | 2.467 | 0.014 |
| Cognitive Learning | Present | 3 | 0.398 | 0.336 | 0.113 | -0.261 | 1.056 | 1.184 | 0.236 |
| Cognitive Learning | Monologue | 16 | 0.237 | 0.146 | 0.021 | -0.049 | 0.524 | 1.624 | 0.104 |
| Cognitive Learning | Personalized | 14 | 0.182 | 0.156 | 0.024 | -0.123 | 0.488 | 1.171 | 0.242 |
| Training Performance | Present | 1 | -0.451 | 0.326 | 0.106 | -1.090 | 0.188 | -1.383 | 0.167 |
| Training Performance | Monologue | 11 | 0.294 | 0.063 | 0.004 | 0.172 | 0.417 | 4.696 | 0.000 |
| Training Performance | Personalized | 8 | 0.206 | 0.095 | 0.009 | 0.019 | 0.393 | 2.162 | 0.031 |
| Transfer Performance | Present | 2 | 0.415 | 0.313 | 0.098 | -0.198 | 1.029 | 1.327 | 0.184 |
| Transfer Performance | Monologue | 18 | 0.226 | 0.109 | 0.012 | 0.013 | 0.438 | 2.078 | 0.038 |
| Transfer Performance | Personalized | 15 | 0.542 | 0.125 | 0.016 | 0.297 | 0.786 | 4.337 | 0.000 |

Hypothesis IVc stated that the relationship between delivery modality and all training evaluation measures would be positive and stronger for agents that utilize facial expressions to help deliver their message than those that do not utilize facial expressions. This hypothesis was not supported for any of the outcome measures. Results are summarized in Table 19.

Table 19: *Hypothesis IVc Overall Results Summary*

| Dependent Variable | Level | *Q*-value | df | *p*-Value | Adj *p* Crit | k |
|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | Overall | 0.741 | 1 | 0.389 | 0.003 | 17 |
| Cognitive Learning | Overall | 0.641 | 1 | 0.423 | 0.002 | 33 |
| Training Performance | Overall | 2.310 | 1 | 0.129 | 0.003 | 20 |
| Transfer Performance | Overall | 8.264 | 1 | 0.004 | 0.001 | 35 |

Given that none of the observed *Q*-values are significant based on the family-wise adjusted *p*-values, observed differences between pedagogical agents in terms of facial expression capabilities are likely due to error as opposed to real differences. These differences are summarized in Table 20.

Table 20: *Facial Expression Results Summary*

| Dependent Variable | Level | k | Point Estimate | Std Err | Variance | Lower Limit | Upper Limit | *Z*-Value | Obs *p*-Value |
|---|---|---|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | No | 2 | 0.041 | 0.279 | 0.078 | -0.506 | 0.588 | 0.146 | 0.884 |
| Post-Training Self-Efficacy | Yes | 15 | 0.291 | 0.083 | 0.007 | 0.129 | 0.454 | 3.511 | 0.000 |
| Cognitive Learning | No | 9 | 0.361 | 0.194 | 0.038 | -0.019 | 0.742 | 1.863 | 0.062 |
| Cognitive Learning | Yes | 24 | 0.180 | 0.116 | 0.014 | -0.048 | 0.409 | 1.548 | 0.122 |
| Training Performance | No | 5 | 0.061 | 0.132 | 0.017 | -0.198 | 0.320 | 0.460 | 0.646 |
| Training Performance | Yes | 15 | 0.283 | 0.062 | 0.004 | 0.162 | 0.403 | 4.587 | 0.000 |
| Transfer Performance | No | 14 | 0.639 | 0.122 | 0.015 | 0.399 | 0.878 | 5.231 | 0.000 |
| Transfer Performance | Yes | 21 | 0.200 | 0.091 | 0.008 | 0.021 | 0.379 | 2.192 | 0.028 |

Hypothesis IVd stated that the relationship between delivery modality and all training evaluation measures would be positive and stronger for agents that

utilize gestures to help deliver their message than those that do not utilize gestures. This hypothesis was not supported for any of the outcome measures. Results are summarized in Table 21.

Table 21: *Hypothesis IVd Overall Results Summary*

| Dependent Variable | Level | *Q*-value | df | *p*-Value | Adj *p* Crit | k |
|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | Overall | 0.490 | 1 | 0.484 | 0.003 | 17 |
| Cognitive Learning | Overall | 1.420 | 1 | 0.233 | 0.002 | 33 |
| Training Performance | Overall | 0.856 | 1 | 0.355 | 0.003 | 20 |
| Transfer Performance | Overall | 2.852 | 1 | 0.091 | 0.001 | 35 |

Given that none of the observed *Q*-values are significant based on the family-wise adjusted *p*-values, observed differences between pedagogical agents in terms of gesture usage are likely due to error as opposed to real differences. These differences are summarized in Table 22.

Table 22: *Gesture Usage Results Summary*

| Dependent Variable | Level | k | Point Estimate | Std Err | Variance | Lower Limit | Upper Limit | *Z*-Value | Obs *p*-Value |
|---|---|---|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | No | 10 | 0.214 | 0.115 | 0.013 | -0.011 | 0.439 | 1.864 | 0.062 |
| Post-Training Self-Efficacy | Yes | 7 | 0.327 | 0.114 | 0.013 | 0.103 | 0.551 | 2.863 | 0.004 |
| Cognitive Learning | No | 13 | 0.080 | 0.160 | 0.026 | -0.234 | 0.393 | 0.499 | 0.618 |
| Cognitive Learning | Yes | 20 | 0.324 | 0.129 | 0.017 | 0.072 | 0.577 | 2.521 | 0.012 |
| Training Performance | No | 10 | 0.180 | 0.087 | 0.008 | 0.009 | 0.350 | 2.067 | 0.039 |
| Training Performance | Yes | 10 | 0.289 | 0.081 | 0.007 | 0.131 | 0.447 | 4.035 | 0.000 |
| Transfer Performance | No | 6 | 0.092 | 0.180 | 0.032 | -0.261 | 0.444 | 0.509 | 0.611 |
| Transfer Performance | Yes | 29 | 0.429 | 0.088 | 0.008 | 0.258 | 0.601 | 4.895 | 0.000 |

Hypothesis V stated that the relationship between support control and all training evaluation measures would be positive and stronger for training scenarios that allow learner control over support delivery than those that require the agent to determine when support is delivered. This hypothesis was not supported for any of the outcome measures. Results are summarized in Table 23.

Table 23: *Hypothesis V Overall Results Summary*

| **Dependent Variable** | **Level** | ***Q*-value** | **df** | ***p*-Value** | **Adj *p* Crit** | **k** |
|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | Overall | 0.156 | 1 | 0.693 | 0.003 | 17 |
| Cognitive Learning | Overall | 2.275 | 1 | 0.132 | 0.002 | 33 |
| Training Performance | Overall | 0.222 | 1 | 0.637 | 0.003 | 20 |
| Transfer Performance | Overall | 0.657 | 1 | 0.417 | 0.001 | 35 |

Given that none of the observed *Q*-values are significant based on the family-wise adjusted *p*-values, observed differences between pedagogical agents in terms of control over support delivery are likely due to error as opposed to real differences. These differences are summarized in Table 24.

Table 24: *Support Delivery Control Results Summary*

| Dependent Variable | Level | k | Point Estimate | Std Err | Variance | Lower Limit | Upper Limit | *Z*-Value | Obs *p*-Value |
|---|---|---|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | Agent | 10 | 0.249 | 0.098 | 0.010 | 0.057 | 0.441 | 2.544 | 0.011 |
| Post-Training Self-Efficacy | Learner | 7 | 0.318 | 0.146 | 0.021 | 0.033 | 0.604 | 2.185 | 0.029 |
| Cognitive Learning | Agent | 23 | 0.137 | 0.115 | 0.013 | -0.089 | 0.363 | 1.192 | 0.233 |
| Cognitive Learning | Learner | 10 | 0.471 | 0.189 | 0.036 | 0.101 | 0.840 | 2.495 | 0.013 |
| Training Performance | Agent | 12 | 0.218 | 0.073 | 0.005 | 0.076 | 0.361 | 2.998 | 0.003 |
| Training Performance | Learner | 8 | 0.279 | 0.107 | 0.011 | 0.070 | 0.488 | 2.611 | 0.009 |
| Transfer Performance | Agent | 24 | 0.323 | 0.092 | 0.008 | 0.143 | 0.504 | 3.512 | 0.000 |
| Transfer Performance | Learner | 11 | 0.463 | 0.145 | 0.021 | 0.178 | 0.747 | 3.189 | 0.001 |

Hypothesis VI stated that the relationship between support timing and all training evaluation measures would be negative and stronger for training scenarios that primarily provide support after learners attempt a new task (delayed) than those that primarily provide support prior to or during (just-in-time) learners attempting a new task. This hypothesis was not supported for any of the outcome measures. Results are summarized in Table 25.

Table 25: *Hypothesis VI Overall Results Summary*

| Dependent Variable | Level | *Q*-value | df | *p*-Value | Adj *p* Crit | k |
|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | Overall | 0.204 | 1 | 0.652 | 0.003 | 17 |
| Cognitive Learning | Overall | 8.824 | 2 | 0.012 | 0.002 | 33 |
| Training Performance | Overall | 6.110 | 2 | 0.047 | 0.003 | 20 |
| Transfer Performance | Overall | 6.688 | 1 | 0.100 | 0.001 | 35 |

Given that none of the observed *Q*-values are significant based on the family-wise adjusted *p*-values, observed differences between pedagogical agents in terms of support timing are likely due to error as opposed to real differences. These differences are summarized in Table 26.

Table 26: *Support Timing Results Summary*

| **Dependent Variable** | **Level** | **k** | **Point Estimate** | **Std Err** | **Variance** | **Lower Limit** | **Upper Limit** | ***Z*-Value** | **Obs *p*-Value** |
|---|---|---|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | Prior to Need | 13 | 0.287 | 0.089 | 0.008 | 0.113 | 0.462 | 3.232 | 0.001 |
| Post-Training Self-Efficacy | Just-In-Time | 4 | 0.191 | 0.194 | 0.038 | -0.189 | -.571 | 0.986 | 0.324 |
| Post-Training Self-Efficacy | Delayed | - | - | - | - | - | - | - | - |
| Cognitive Learning | Prior to Need | 19 | 0.024 | 0.121 | 0.015 | -0.214 | 0.261 | 0.194 | 0.846 |
| Cognitive Learning | Just-In-Time | 13 | 0.575 | 0.151 | 0.023 | 0.280 | 0.870 | 3.823 | 0.000 |
| Cognitive Learning | Delayed | 1 | -0.182 | 0.509 | 0.259 | -1.179 | 0.815 | -0.358 | 0.720 |
| Training Performance | Prior to Need | 11 | 0.325 | 0.060 | 0.004 | 0.207 | 0.443 | 5.406 | 0.000 |
| Training Performance | Just-In-Time | 7 | 0.046 | 0.110 | 0.012 | -0.169 | 0.262 | 0.423 | 0.672 |
| Training Performance | Delayed | 2 | -0.029 | 0.264 | 0.070 | -0.547 | 0.488 | -0.111 | 0.912 |
| Transfer Performance | Prior to Need | 19 | 0.194 | 0.098 | 0.010 | 0.002 | 0.387 | 1.979 | 0.048 |
| Transfer Performance | Just-In-Time | 16 | 0.584 | 0.115 | 0.013 | 0.360 | 0.809 | 5.102 | 0.000 |
| Transfer Performance | Delayed | - | - | - | - | - | - | - | - |

Research Question I asked, does focusing on any of the four objects of support (i.e., content, problem-solving, meta-cognition, technology) result in improved learning outcomes more so than focusing on other objects of support? Overall results indicate that no particular object of support results in better learning outcomes than other objects of support. Results are summarized in Table 27.

Table 27: *Research Question I Overall Results Summary*

| Dependent Variable | Level | *Q*-value | df | *p*-Value | Adj *p* Crit | k |
|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | Overall | 0.723 | 3 | 0.868 | 0.003 | 17 |
| Cognitive Learning | Overall | 1.554 | 3 | 0.670 | 0.002 | 33 |
| Training Performance | Overall | 0.312 | 2 | 0.856 | 0.003 | 20 |
| Transfer Performance | Overall | 7.622 | 3 | 0.055 | 0.001 | 35 |

Given that none of the observed *Q*-values are significant based on the family-wise adjusted *p*-values, observed differences between pedagogical agents in terms of object of support are likely due to error as opposed to real differences. These differences are summarized in Table 28.

The implications of the results described above on theory and practice will be discussed in the next section. Additional thoughts, questions, concerns, and explanations for the expected and unexpected results will also be brought up and addressed.

Table 28: *Object of Support Results Summary*

| Dependent Variable | Level | k | Point Estimate | Std Err | Variance | Lower Limit | Upper Limit | *Z*-Value | Obs *p*-Value |
|---|---|---|---|---|---|---|---|---|---|
| Post-Training Self-Efficacy | Cannot Determine | 2 | 0.240 | 0.210 | 0.044 | -0.171 | 0.652 | 1.146 | 0.252 |
| Post-Training Self-Efficacy | Content | 10 | 0.299 | 0.117 | 0.014 | 0.069 | 0.529 | 2.551 | 0.011 |
| Post-Training Self-Efficacy | Problem Solving | 2 | 0.041 | 0.297 | 0.088 | -0.542 | 0.623 | 0.137 | 0.891 |
| Post-Training Self-Efficacy | Meta-Cognition | 3 | 0.314 | 0.205 | 0.042 | -0.088 | 0.717 | 1.529 | 0.126 |
| Cognitive Learning | Cannot Determine | 2 | -0.124 | 0.391 | 0.153 | -0.890 | 0.642 | -0.316 | 0.752 |
| Cognitive Learning | Content | 24 | 0.220 | 0.119 | 0.014 | -0.013 | 0.452 | 1.851 | 0.064 |
| Cognitive Learning | Problem Solving | 6 | 0.336 | 0.247 | 0.061 | -0.148 | 0.820 | 1.361 | 0.173 |
| Cognitive Learning | Meta-Cognition | 1 | 0.688 | 0.621 | 0.386 | -0.530 | 1.906 | 1.107 | 0.268 |
| Training Performance | Cannot Determine | 1 | 0.295 | 0.275 | 0.076 | -0.244 | 0.833 | 1.073 | 0.283 |
| Training Performance | Content | 11 | 0.252 | 0.075 | 0.006 | 0.105 | 0.399 | 3.360 | 0.001 |
| Training Performance | Problem Solving | 8 | 0.179 | 0.121 | 0.015 | -0.059 | 0.417 | 1.474 | 0.140 |
| Training Performance | Meta-Cognition | - | - | - | - | - | - | - | - |
| Transfer Performance | Cannot Determine | 3 | -0.098 | 0.230 | 0.053 | -0.548 | 0.353 | -0.425 | 0.671 |
| Transfer Performance | Content | 22 | 0.327 | 0.096 | 0.009 | 0.139 | 0.514 | 3.414 | 0.001 |
| Transfer Performance | Problem Solving | 7 | 0.634 | 0.176 | 0.031 | 0.289 | 0.979 | 3.601 | 0.000 |
| Transfer Performance | Meta-Cognition | 3 | 0.658 | 0.283 | 0.080 | 0.103 | 1.214 | 2.323 | 0.020 |

Given the results presented above, the next section will discuss the strengths of the present study, its weaknesses, and the theoretical and practical implications that can be drawn from the data.

# Discussion

## Study Strengths

Though relatively few of the proposed hypotheses achieved statistical significance, there were many strengths of the present study worth highlighting. The first of these strengths is the sampling methodology. A variety of sources were used to attain relevant research studies, including multiple online database searches (utilizing intentionally broad search terms), combing recently published journals for relevant articles, and contacting researchers who are or have been active in the pedagogical agent domain (professionally and academically) to acquire any existing unpublished studies. These efforts resulted in nearly 5,000 articles to review for potential inclusion in the present study. While a large proportion of these studies were irrelevant to the goals of this study, the wide array of results is a testament to the comprehensiveness of the search process.

Relatedly, this study contained more than double the percentage of unpublished studies commonly found in meta-analyses (Borenstein, et al., 2009). Despite the low level of correspondence from researchers involved in this area of study, the nature of research on this topic seems to have made unpublished studies more accessible than they might be in other areas of research. As a very digital-oriented research topic, researchers in this area seemed quite ready to share doctoral dissertations, conference presentations, and other unpublished manuscripts via personal and institutional websites. Interest in and access to digital knowledge sharing outlets may have led to increased electronic availability

of these documents compared to documents in other, less technology-centric areas of research.

As a result of this comprehensive article search, publication bias did not seem to be of statistical concern. In all cases, Orwin's Fail-Safe N indicated that at least double the number of studies available would have been required to decrease the observed effect size for each dependent variable below a "trivial" level.

Also with regard to the statistical methods of this study, care was taken to perform a Type I error correction when considering the statistical significance of each analysis. As a complex study with multiple hypotheses, multiple levels of each independent variable, and multiple dependent variables, the likelihood of capitalizing on chance to achieve statistical significance was very high. As Pigott and Polanin (2014) discussed, meta-analytic researchers often fail to address this consideration. As such, the decision to address this issue and implement a relatively stringent Type I error correction should be considered a strength of this study in particular.

A final strength worth noting is the breadth of professional and academic domains that contributed articles to this study. The final data set included studies conducted by Cognitive Psychologists, Industrial/Organizational Psychologists, business researchers, training specialists, educators, researchers in the domain of human-computer interaction, and many others. This helped to increase the variety of theoretical bases considered, study designs implemented, and analytic methods

employed. In turn, this helps to increase the representativeness of the sample and generalizability of the results.

**Study Weaknesses**

Despite the various strengths of this study, there are a few weaknesses worth mentioning as well. One of the most obvious limitation is that the final sample of studies was relatively homogeneous in terms of the participants used. The samples were coded as "students" in a "lab" setting in 57 of the 59 unique samples collected from the 41 studies. This means that 96.6% of the participants were adult (typically college) students who would probably be considered a "convenience sample". While offering a substantial amount of internal control over the testing conditions, these samples likely restricted the age range and range of prior experiences, knowledge, skills, and abilities the participants brought to the individual studies, and also limited the authenticity of experiencing training under "real world" conditions (with "real world" implications). As such, this sample of studies likely limits the generalizability of the findings.

Another potential source of concern is the lack of representation of highly-regarded journals from which the studies were drawn. While some notoriously rigorous journals do appear in the list from various domains (e.g., Applied Cognitive Psychology, Journal of Applied Psychology, Journal of Educational Psychology, Computers in Human Behavior, etc.), most of the included studies originated in journals with much shorter publication histories and less prestigious reputations. While most of the articles did undergo a peer review process and appear to have taken measures to ensure proper study design, analysis, and

reporting, it is unclear what was the true academic rigor of many of the studies in this meta-analysis. While multiple studies would need to be similarly skewed in the same direction in order to influence the results of the present meta-analysis, given that this study has no way to measure this possible impact it is worth highlighting as a potential concern.

From a statistical perspective, while care was taken to utilize an appropriately strict *p*-critical value to minimize the risk of Type I error, the lack of agreement on what is the most appropriate adjustment formula implies that the adjustment used in this study could realistically be too strict. This potential is evidenced by the number of observed *p*-values that surpassed traditional *p*-critical values of 0.05 or 0.01, but did not surpass the various adjusted *p*-critical values. As such, "nearly significant" and "marginally significant" results are also elaborated below.

**Hypothesis Testing**

In line with the predictions of the first hypothesis, it appears that making human-like instructional characters increasingly lifelike may not always result in optimal learning outcomes. This is evidenced by the fact that transfer task performance was significantly higher for participants who learned from human-like agents who were moderately iconic (neither too cartoon-like nor too realistic). These results support the concept behind Mori's (1970) Uncanny Valley that it may be possible to design a robot or instructional character (that will never be completely human) to be a little too realistic such that people begin to react negatively to it and become distracted from the task at hand (in this case,

learning). However, care should also be taken to design pedagogical agents that are not too cartoon-like, but instead exhibit some degree of fine lines, shading, detail, and realistic proportions to help generate the social connection and affinity required to engage learners, meet their expectations of the training program, and elicit learning.

Practically speaking, this is good news for companies seeking to use pedagogical agents as part of their comprehensive training programs. It indicates that (at least as far as human-like trainers are concerned), efforts and special software to make training agents appear as human-like as possible may not be necessary to help ensure the KSAs being taught in the training program successfully translate to performance on the job. Moderately realistic trainers may be interpreted as being real enough such that learners identify with them and learn the information more deeply, resulting in a better ability to apply the information learned to novel tasks. Conversely, moderately realistic training agents are not so realistic as to set unachievably high expectations for learners, only to fall short of them and/or distract learners from the task at hand (such that learning is hindered).

With regard to the iconicity of non-human-like pedagogical agents, the article search uncovered relatively few non-human agents to compare. No analysis could be conducted for post-training self-efficacy or cognitive learning. Additionally, no articles qualified for the non-human-like/low iconicity (high realism) category. As such, the only comparisons that could be drawn were across agents categorized as moderate and high on iconicity (low realism). Though the

results of the analysis using the adjusted *p*-critical values were not statistically significant, the observed *p*-value for transfer task performance overall would be significant at most conventional *p*-critical values (e.g., < 0.05 and < 0.01). Additionally, it would have been significant at the Borenstein, et al. (2009) recommended cut-off of 0.01. As such it may be worth considering that the high-iconicity, non-human-like (cartoon-like) training agents produced significantly higher transfer task scores than the slightly more lifelike agents.

This could be a result of increased perceptions of congruity on the part of the learners, such that the wizards, genies, robots, and bugs used to deliver these training programs created less cognitive dissonance, confusion, and distraction as cartoon characters than they did as slightly more realistic depictions. While this explanation would be contrary to the hypotheses, the results would be consistent with cognitive psychology principles presented earlier suggesting that when pedagogical agent depictions create unrealistic expectations for learners and then fail to deliver on those expectations (in terms of serving as social beings to interact with and relate to), the trainees will be distracted from the task and hand, which will subsequently result in poorer learning outcomes. It could be that these fictional characters (as non-human-like beings) are best (and ideally) represented as less realistic.

The implications of these results for anyone seeking to implement pedagogical agents would be that, again, few efforts should be made to make these characters overly lifelike and realistic. It may resonate better with learners to present inanimate or non-human characters as close to prototypical cartoon

characters as possible. The adults in the studies represented here may have grown up learning lessons and knowledge from any number of low-fidelity cartoon characters, and thus have little trouble focusing on the non-human, cartoon-like characters used in these studies to train various knowledge and task-related skills.

Taken together, the results of Hypotheses Ia and Ib would suggest to practitioners seeking to develop training programs to use human-like pedagogical agents whenever possible, but to de-emphasize making them as realistic as possible, instead opting for a moderate level of realism. The results of the present study suggest this approach could lead to optimal transfer performance, which may in turn lead to measurable organizational results. Since the results for non-human-like trainers were less convincing, practitioners may be best served to avoid the use of non-human-like characters unless they are particularly relevant to the content of the training. If these characters are to be used in training, it may be best to present them as cartoon-like to potentially increase transfer task performance.

Hypotheses II and III were very closely related, and dealt specifically with the instructional roles and modalities that can be programmed into pedagogical agents. Referring again to Table 1, it is easy to see the overlap between these two constructs as presented in Clarebout, et al. (2002). Instead of discussing the instructional *roles* as phrased in their framework (“Supplanting”, “Scaffolding”, “Demonstrating”, “Modeling”, “Coaching”, and “Testing”) it may be easier to think about them as actual roles (“Supplanter”, “Scaffolder”, “Demonstrator”, “Modeler”, “Coach”, and “Tester”). The instructional *modalities*, then, are the

actions performed by agents in their roles. For example, Supplanters merely Execute tasks for learners to observe, whereas Modelers Show and Explain procedures and tasks, while Coaches Explain and Question, etc.

Given the overlap of these distinct but related constructs, it is no surprise that the overall results for these hypotheses were so similar. In both cases, the results for training performance and transfer performance were not significant at the adjusted *p*-critical values, but the observed *p*-values were near or lower than the standard critical value of $p < 0.05$ making it worth considering their practical significance.

Taken together, the overall results suggest there may be a directional relationship between pedagogical agent behaviors (that elicit varying levels of cognitive processing) and subsequent training outcomes. However, the detailed results for these hypotheses are a little more convoluted. For example, Executing and Showing instructional modalities (which elicit shallow cognitive processing) seem to be related to increased performance on training tasks. Further, (except for 2 studies where the modality could not be determined) all modalities (regardless of the depth of processing elicited) were significantly and positively related to transfer task performance.

Additionally, when Supplanters (an instructional role that elicits a shallow depth of processing) delivered the training sessions, it resulted in significantly higher scores on training tasks, while Supplanters and Scaffolders (again, eliciting shallow processing) resulted in significantly higher transfer scores than other roles eliciting deeper processing. The exception to this is Testers (who are supposed to

elicit the deepest level of processing), who also elicited significantly higher transfer scores. These results do not support the Depth of Processing hypotheses.

From a theoretical standpoint, the Depth of Processing results may indicate that the level to which trainees engage with the material when working with pedagogical agents may not be the best predictor of training outcomes. Simply presenting well-planned and organized content with the help of an interesting, engaging, digital trainer may be enough to transmit the required knowledge, skills, and abilities. This would align with Baylor (2000) and her views of what makes pedagogical agents effective. As noted above, she stated that, for pedagogical agents to be effective mentors and trainers, they must display "regulated intelligence, exhibit some persona, and display pedagogical control". In addition, she described the functional elements of the "agent metaphor", which is simply the visible presence of an agent in a learning program. According to this view, the mere presence of a pedagogical agent (a social being) who delivers a competent lecture may be effective over other training programs that do not meet these criteria.

This more parsimonious view of the relationship between agents and their impacts on learning would be supported by research on Social Agency Theory and the presence/absence of pedagogical agents in training programs. "Social agency theory" states that "social cues in a multimedia message can prime the social conversation schema in learners" (Louwerse, Graesser, Lu, & Mitchell, 2005). Applied to pedagogical agents, this means that seeing an anthropomorphized character in a training setting might trigger responses typical

for human to human social interactions. These responses in a learning setting may make learners more likely to pay attention and engage with the material as they would in a variety of human-human training scenarios.

Early work related to this theory includes a meta-analysis by Kim and Ryu (2003) looking at 28 different pedagogical agent studies. According to the authors, the mere presence of a pedagogical agent in a training program resulted in greatly improved retention and transfer test scores. The authors suggest that these results may be due to the motivational effects of being in the presence of a social being, such that people simply want to perform better in front of a pedagogical agent than they do when they perceive themselves to be alone during computer-mediated training. In sum, the seemingly mixed and convoluted results for Hypotheses II and III could indicate that the behaviors pedagogical agents perform may be less important to predicting training performance than the way the agent looks, moves, and is perceived by the learners. As such, this puts the onus on training practitioners to develop high quality content delivered by pedagogical agents who meet certain superficial appearance criteria, rather than crafting complex behavioral algorithms to elicit one instructional role or modality over another.

Hypotheses IVa - IVd shared a common theme of exploring the delivery mechanisms that can be programmed into pedagogical agents. Hypothesis IVa argued that agents who present content using speech instead of text as the primary delivery mechanism would result in the most optimal training outcomes. The three levels of the independent variable were speech, text, or merely “present”.

Agents who were coded as “present” contributed to the content delivery by moving throughout the training environment, directing learner attention to various images, diagrams, and other information while a narration delivered the content. However, it was unclear in the training if the agent was intended to be the speaker.

The overall test for significance for Hypothesis IVa met the adjusted *p*-critical value of 0.001 for transfer performance. A key detail to note is that 32 of the 35 studies included in this analysis utilized speech as the primary delivery mechanism. Looking at the subgroup results, speech was significant with regard to generating high transfer task scores. Though text was also identified as being positively and significantly related to transfer scores, this result is based on only one study.

Given the available data, it appears that Hypothesis IVa was at least partially supported. Pedagogical agents that deliver content primarily through speech may more reliably elicit successful transfer task performance over agents who rely solely on their physical movements to communicate content to learners. This aligns with the cognitive psychological theories mentioned above that state agents who present information via speech do not overly tax any one of the dual processing channels (Baddeley, 1992, 1998; Clark & Paivio, 1991; Paivio, 1986), capitalize on the modality effect (Ginns, 2005), and fit well within the Cognitive Theory of Multimedia Learning (Mayer, 2001). Until more studies are conducted and more comparisons can be made between the speech and text conditions, practitioners may best be served by creating pedagogical agents that deliver

content via speech as it is unlikely to violate various learning and cognitive psychology principles. Delivery via speech also likely contributes to perceptions of the agents as social beings (and not abnormal and distracting), which is a principle that has been critical to the results seen in other hypotheses in this study.

Hypothesis IVb stated that the relationship between agents who used personalized messages would be more beneficial to training outcomes than agents who simply delivered a lecture using a more monologue style. The overall results for this hypothesis were not significant for any of the training outcomes measured. Personalized delivery was supposed to engage learners, make the lesson seem more personal, and encourage participation, potentially leading to deeper information processing and subsequent learning. Additionally, interaction and feedback exchanges between agents and learners were supposed to contribute to the social nature of a training scenario, which was also supposed to facilitate learning.

One explanation for these results could be related to the "social agency" theories presented above that state that the agents coded in this study may have elicited enough interest and engagement as social beings on their own, and there were no incremental benefits of utilizing social feedback and response techniques. It could be that pedagogical agents generate sufficient interest in the learning task, and that their efforts to connect to learners individually could be considered behaviors that violate learner expectations of the agents' pedagogical abilities, and thus become detrimental (or at least) distracting, and lead to the observed null effects.

Hypothesis IVc suggested that agents who are programmed with advanced facial expressiveness will produce better learning outcomes than those who are less expressive. This hypothesis was not supported. One potential and practical reason for this is that opportunities to observe detailed agent expressions may be limited during a training session. Agents are frequently a fraction of the size screen on which they are observed, they do not always face directly at the learner, and their small faces may be blurred by grainy computer monitor resolutions. Ultimately, many studies failed to conduct manipulation checks to ensure their respondents were reacting as expected to the agents they worked with, and the present study did not code for the quality or extent to which the agent was expressive, so it is very possible that participants had a difficult time observing the programmed emotions, reactions, and expressions.

Additionally, unless substantial time and effort is invested in developing appropriate and detailed agent reactions and emotions, many out-of-the-box agent development software programs may not deliver 100% accurate or appropriate expressions throughout the entire training session. Many of these programs can broadly apply common emotions such as “happy”, “sad”, or “angry”, but without complex programming, expressions rarely adjust automatically to fit the content being delivered. For example, an agent pre-programmed to display happiness may appear to be inappropriately or unrealistically happy throughout the entire training, which could diminish the impact of those expressions.

Recommendations for practitioners based on these results would be to avoid expending effort developing elaborate facial expressiveness, especially if

the agent's face is not the most prominent focal point of the training. While it is still important to create agents that are likeable and relatable to learners (to trigger the appropriate social connections), simply avoiding expressions that could be interpreted negatively by learners may be sufficient for eliciting the desired training outcomes. Additionally, agents should be pilot tested prior to training to ensure participants perceive the agents as expected. These measures should also be used during or immediately after the training to gauge their impact during the training as well.

Hypothesis IVd was the final delivery modality hypothesis presented, and it suggested that agents who seamlessly incorporate deliberate body movements and gestures into the training program would be more effective trainers than those who do not incorporate those gestures. This hypothesis was not supported. Similar to the facial expressions discussed above, programming intricate, natural, and well-timed gestures often requires highly advanced and technical knowledge and skill. While many of the programs used to develop these agents incorporate features that facilitate natural, fluid movements, these movements may be too general or not direct enough with regard to directing learner attention to specific elements of the training program. For example, even a task as simple as pointing to a piece of information on the other side of a computer monitor requires, at the very most, complex technical programming of the finger, hand, elbow, shoulder, and torso, and at the very least, it requires timing the gesture according to the content delivery so as to not make the gesture too slow, too fast, or too errant.

Given that neither the quality nor the amount of gestures were coded for this study (and rarely discussed within the original articles), it is difficult to determine how well the authors of the original studies programmed their pedagogical agents to execute these tasks. The imprecision or potential awkwardness of the complex movements incorporated into various training programs could lead to distraction from the learning task at hand, or at the very least it could contribute to decreasing the realism or believability of the trainer (which, as shown above, is relevant to the success of the training program).

As such, when designing pedagogical agent training programs, it may be advisable to avoid programming complex gestures or forcing the agents to intricately refer to very specific elements of the training. When done incorrectly, it may appear unnatural or forced. Provided the timing is correct, programming simple movements like weight shifting, shoulder shrugs, head tilts, and subtle hand gestures as the agent speaks may help to foster the illusion of life, but as the state of the technology stands, it does not appear that attempting to program more nuanced movements results in improved learning outcomes.

In looking at the results of Hypothesis V, it does not appear that giving learners control over the delivery of support leads to improved learning outcomes as hypothesized. However, this is one potential area of the study where a more nuanced operationalization of the independent variable could have been of use. For example, learner control could have been as simple as a pause and rewind function in one study, whereas another study might allow learners to click a "help" button that prompts the pedagogical agent to elaborate on a specific topic,

whereas a third study may program a complex network of topics, examples, elaborations, and self-tests that the learners navigate on their own to learn the material being trained. This is a very broad range in which "learner control" can be defined, so from a theoretical perspective, this definition does little to advance the utility of the Clarebout, et al. (2002) support typology used in this study, and from a practical perspective it allows us to make few recommendations with regard to who should dictate the pace and elaboration of information in a training program. As it stands, it may not hurt to program training programs with pause, rewind, and/or various tools to allow for information elaboration, but the current data does not make a strong case for this being "better" or beneficial for training outcomes.

Another aspect of the Clarebout, et al. (2002) support typology examined in this study is the idea that support for learning can be delivered by a pedagogical agent at various times relative to the learner's need to apply that knowledge, and this idea was explored in Hypothesis VI, however, it was not supported. From a statistical standpoint, across 4 dependent variables, only 3 studies qualified as delivering delayed support to learners. This leaves the primary comparison to be between support delivered "prior to need" or "just-in-time". Many of the agents coded as delivering support prior to needing the information also utilized a monologue, or lecture style approach, whereas the just-in-time support deliverers typically offered their assistance during times when learners were practicing a task or working through a problem and needed a little help. As such, the lack of significant results could be due to overlap with the instructional modalities

discussed above. Agents presenting information prior to need and utilizing a lecture-style approach are likely utilizing a Supplanting or Scaffolding instructional modality, while agents providing help during a practice task are likely utilizing a Testing or Coaching instructional modality. As was seen earlier, the results for instructional modality did not suggest much support for eliciting different levels of cognitive processing (and subsequent levels of training outcome performance). Given the high overlap between the coding of these constructs and the lack of data available for delayed support delivery, it becomes clearer why this hypothesis may have failed to reach significance.

As it was discussed in the Clarebout et al., (2002) support typology, there was little reason to believe any individual object of support (i.e., area of the training toward which the agent directs and focuses its assistance; content, problem-solving, meta-cognition, technology) would result in improved learning outcomes than any other. The rationale was that any support provided above and beyond the base presentation of content should aid the learner in his/her pursuit of knowledge about the topic at hand, and Research Question I sought to explore this possibility.

Looking across independent variable – dependent variable comparisons, the majority of agents that were coded focused their support on elaboration of content and aiding in problem solving as opposed to helping learners manage meta-cognitive processes. None of the agents focused their support on the technology or tools available to learners. Given that none of the overall relationships proved to be statistically significant, it may be said that this study

offers no evidence that focusing support on any of the four potential objects of support offers benefits over the others. As such, the best practice for pedagogical agents in applied settings may be to pilot test the training to determine the areas where trainees are most likely to get stuck. At those points, the agents can be programmed to offer whatever support may be necessary. For example, if certain content proves to be exceedingly difficult for most learners, content support can be offered. Alternatively, if the training program is quite long or has many interconnected parts, offering meta-cognitive guidance to help learners navigate the information may be beneficial.

**Future Research**

Despite the mixed results of this study, the Integrated Model of Training Evaluation and Effectiveness (Alvarez, et al., 2004) was a useful overarching framework with which to analyze these pedagogical agents. It provided a convenient and logical guideline for mapping effectiveness criteria (i.e., post-training self-efficacy, cognitive learning, training performance, and transfer performance) to the training characteristics of interest (specifically, the various pedagogical agent appearances, behaviors, and personas). However, this study primarily evaluated the "middle" of this model by focusing on training characteristics and the desired changes in learners we would hope to see. What the present study did not examine was any individual/trainee characteristics that can impact the results of a training session. Individuals and the unique knowledge, skills, abilities, perceptions, and reactions they bring to a training session can impact any part of a training process, from before the training starts until long

after the training ends. The scope of this study did not include any trainee reactions to the training content or design. Future research may be able to use trainee reaction information as a covariate or moderator in similar analyses as those presented here. Filling in this information may help to uncover nuances and relationships not detected with the present study's design and methodology.

Relatedly, many of the mixed or non-significant results observed in this study may be the result of aptitude-treatment interaction effects. Aptitude-treatment interaction effects refer to the idea that some types of training (treatment) may be more or less effective for certain people depending on their individual abilities (aptitude) (Cronbach & Snow, 1977). This concept is illustrated in the training effectiveness section of the IMTEE (Alvarez et al., 2004) by the fact that individual characteristics and training characteristics can simultaneously influence any and all of the training evaluation measures (the study's dependent variables). Given that the current study did not account for learner abilities with regard to the pedagogical agents' designs and behaviors, an opportunity exists for future research to explore these relationships.

Toward this end, the Comprehensive Meta-Analysis software program would allow for these types of analyses via a meta-regression. Similar to a "standard" regression (where respondents are the unit of measurement), a meta-regression performs similar analyses using individual studies as the unit of measurement. As such, individual-level and training-level characteristics could be entered as covariates to measure their effects (unique and interactive) on the dependent variables (in this case our training evaluation measures). As efforts to

customize, personalize, and bring learning down to the level of the individual learner increase through the use of pedagogical agents, these aptitude-treatment effects will need to be examined.

On the opposite end of the training evaluation spectrum, the present study did not test the ultimate organizational evaluation criteria of "results". These are the bottom line impacts that organizations experience as outcomes of training programs. However, with less than 4% of our total sample representing research conducted in actual organizations (instead of in a lab), data on these results would be very difficult to come by. If future research can begin to fill in this gap of the impact of pedagogical agents in applied settings, we may be better able to answer the question of what role pedagogical agents serve in the macro world of "employee training".

In terms of future methodologies, it may be interesting to examine pedagogical agents using similar studies, similar frameworks, and similar logic as the present study, but with more granular and specific agent coding practices. This methodology would only be possible by attaining the actual training materials from the original researchers and performing much more detailed coding of the agent appearances and behaviors. Ultimately, the level of detail researchers are able and willing to provide in a journal article is much more superficial than being able to see the agents "in action". It would be interesting to see how specific the pedagogical agent design recommendations could become, and how results of that study would compare to the results of this study.

Another potential change for future research could be refinement to the operationalization of "iconicity". It is possible that the definitions used in this study may not be granular enough. Specifically, the Uncanny Valley (Mori, 1970) is portrayed as a continuous curve, but iconicity in this study was necessarily coded as high/medium/low due to sample size concerns. As some of the more promising results that came out of this study, this area may be worth further exploration to discover exactly where the Uncanny Valley "drop off" is, and what specific characteristics do and do not push agents over the edge from helpful to detrimental to learning. This knowledge could offer many more specifics regarding the optimal design of pedagogical agents.

A methodological issue related to the last point is that there is potential for range restriction in the present study with regard to the Uncanny Valley. The present study did not examine industrial, lifeless robots on the lower end of the Uncanny Valley spectrum, nor did this study examine high-end, physical representations of human trainers. Research delving into the training potential of these extreme examples may help to paint the full picture of agent relations to humans and the transfer of KSAs.

This meta-analysis explored the various ways in which pedagogical agents can differ in terms of their appearance, pedagogical behaviors, and social behaviors, and the impact these differences can have on various learning outcomes. While the purpose of this study was largely to help guide future pedagogical agent training design, the question still exists as to whether and in what situations pedagogical agents are the right training solution to begin with.

Alternatives include everything from written manuals to audio files to training videos to human trainers. More research is necessary to compare agent conditions to these (and other) no-agent conditions in various training settings to help discern when and with whom pedagogical agents are most useful and when other training alternatives are more effective.

Finally, a meta-analysis by Schroeder, Adesope, and Gilbert (2013) found that pedagogical agents were more effective training agents for learners in grades K-12 than for post-secondary school learners. The effects observed in the present study, therefore, could be stronger for younger learners than the adults sampled across studies in this meta-analysis. It would be interesting to see how the methodology and framework explored in this study would hold across studies testing younger samples of learners. Relatedly, future work could seek to explore how, when, and why this difference develops across learners of different ages. Work in this area could lead to refinements to pedagogical agent training that caters to trainees depending on their age and related information processing abilities and preferences.

## References

* *Indicates articles coded for analysis*

Aguinis, H., & Kraiger, K. (2009). Benefits of training and development for individuals and teams, organizations, and society. *Annual Review of Psychology, 60,* 451-474.

Alvarez, K., Salas, E., & Garofano, C.M. (2004). An integrated model of training evaluation and effectiveness. *Human Resource Development Review, 3*(4), 385-416.

Andre, E., Rist, T., & Muller, J. (1999). Employing AI methods to control the behavior of animated interface agents. *Applied Artificial Intelligence, 13,* 415-448.

Arthur, W., Bennett, W., Edens, P.S., & Bell, S.T. (2003). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology, 88*(2), 234-245.

ASTD. (2013). 2013 State of the Industry Report. Retrieved January 1, 2014, from http://store.astd.org/Default.aspx?tabid=167&ProductId=25317

*Atkinson, R. K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology, 94*(2), 416.

Baddeley, A. (1992). Working memory. *Science, 255*(5044), 556-559.

Baddeley, A. (1998). Recent developments in working memory. *Current Opinion in Neurobiology, 8,* 234-238.

Bates, J. (1994). The role of emotion in believable agents. *Communications of the ACM, Special Issue on Agents*, 1-6.

Baylor, A.L. (2000). Beyond butlers: Intelligent agents as mentors. *Journal of Educational Computing Research, 22*(4), 373-382.

*Baylor, A.L, & Kim, Y. (2004). *Pedagogical Agent Design: The Impact of Agent Realism, Gender, Ethnicity, and Instructional Role*. Paper presented at the Intelligent Tutoring Systems, Maceio, Alagoas, Brazil. (Vol. 3220, 592-603).

*Baylor, A.L, & Kim, Y. (2005). Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education, 15*, 95-115.

*Baylor, A.L, & Kim, Y., Shen, E. (2007). Pedagogical agents as learning companions: The impact of agent emotion and gender. *Journal of Computer Assisted Learning, 23*, .220–234

Becker, B.E., & Huselid, M.A. (1998). High performance work systems and firm performance: A synthesis of research and managerial implications. *Research in Personnel and Human Resource Management, 16,* 53-101.

Bem, D., Tressoldi, P., Rabeyron, T., & Duggan, M. (2015). Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events. *F1000Research*, 4.

Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Chichester, UK: John Wiley & Sons, Ltd.

Broadbent, D. E. (1958). *Perception and communication*. New York: Oxford University Press.

Burke, L.A., & Hutchins, H.M. (2008). A study of best practices in training transfer and proposed model of transfer. *Human Resource Develoment Quarterly, 19*(2), 107-128.

Cassell, J., & Thorisson, K.R. (1999). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence, 13*, 519-538.

*Choi, S., & Clark, R. E. (2006). Cognitive and affective benefits of an animated pedagogical agent for learning English as a second language. *Journal of Educational Computing Research, 34*(4), 441-466.

*Clarebout, G., & Elen, J. (2005). A pedagogical agent's impact in an open learning environment. In *CELDA* (415-418).

Clarebout, G., Elen, J., Johnson, W. L., & Shaw, E. (2002). Animated pedagogical agents: An opportunity to be grasped? *Journal of Educational Multimedia and Hypermedia, 11*(3), 267-286.

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology review, 3*(3), 149-210.

Colquitt, J.A., LePine, J.A., & Noe, R.A. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology, 85*(5), 678-707.

Combs, J., Liu, Y., Hall, A., & Ketchen, D. (2006). How much do high-performance work practices matter? A meta-analysis of their effects on organizational performance. *Personnel Psychology, 59*(3), 501-528.

*Craig, S. D., Driscoll, D. M., & Gholson, B. (2004). Constructing knowledge from dialog in an intelligent tutoring system: Interactive learning, vicarious learning, and pedagogical agents. *Journal of Educational Multimedia and Hypermedia*, *13*(2), 163.

*Craig, S. D., & Gholson, B. (2002). Does an agent matter? The effects of animated pedagogical agents on multimedia environments. In P. Barker & S. Rebelsky (Eds.), *Proceedings of ED-MEDIA 2002--World Conference on Educational Multimedia, Hypermedia & Telecommunications* (357-362). Denver, Colorado, USA: Association for the Advancement of Computing in Education (AACE). Retrieved September 9, 2017 from https://www.learntechlib.org/p/9301/.

*Craig, S. D., Gholson, B., & Driscoll, D. M. (2002). Animated pedagogical agents in multimedia educational environments: Effects of agent properties, picture features and redundancy. *Journal of Educational Psychology*, *94*(2), 428.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Educational Psychology: General, 104*(3), 268-294.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. Oxford, England: Irvington.

*Dirkin, K. H., Mishra, P., & Altermatt, E. (2005). All or nothing: Levels of sociability of a pedagogical software agent and its impact on student perceptions and learning. *Journal of Educational Multimedia and Hypermedia, 14*(2), 113.

*Domagk, S. (2010). Do pedagogical agents facilitate learner motivation and learning outcomes?. *Journal of Media Psychology*. *22*, 84-97.

*Dunsworth, Q., & Atkinson, R. K. (2007). Fostering multimedia learning of science: Exploring the role of an animated agent's image. *Computers & Education*, *49*(3), 677-690.

Duval, S. J., & Tweedie, R. L. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*, 89 –98.

Duval, S. J., & Tweedie, R. L. (2000b). Trim and fill: A simple funnelplot-based method of testing and adjusting for publication bias in metaanalysis. *Biometrics, 56*(2), 455– 463.

Elen, J. (1995). *Blocks on the road to instructional design prescriptions: A methodology for ID-research exemplified*. Leuven, Belgium: Leuven University Press.

Evers, M., & Nijholt, A. (2000). Jacob- An animated instruction agent in virtual reality *Advances in Multimodal Interfaces- ICMI* (pp. 526-533): Springer Berlin Heidelberg.

Feldman, J. M. (1981). Beyond Attribution Theory: Cognitive Processes in Performance Appraisal. *Journal of Applied Psychology*, *66*(2), 127-148.

*Fujimoto, T. (2010). *Story-based pedagogical agents: A scaffolding design approach for the process of historical inquiry in a web-based self-learning environment*. The Pennsylvania State University. ProQuest Dissertations Publishing, 2010. 3420146.

Ganeshan, R., Johnson, W. L., Shaw, E., & Wood, B.P. . (2000). Tutoring diagnostic problem solving. *Inteligent Tutoring Syatems,* 33-42.

Garner, R., Gillingham, M. G., & White, C. S. (1989). Effects of seductive details on macroprocessing and microprocessing in adults and children. *Cognition and Instruction, 6*, 41–57.

Ginns, P. (2005). Meta-analysis of the modality effect. *Learning and Instruction, 15*, 313-331.

Goldstein, I.L., & Ford, J.K. (2002). *Training in Organizations* (4th ed.). Belmont, CA: Wadsworth.

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., & Kreuz, R. (1999). AutoTutor: A simulation of a human tutor. *Journal of Cognitive Systems Research, 1*, 35-51.

Gregorie, J. P., Zettlemoyer, L. S., & Lester, J. C. (1999). *Detecting and correcting misconceptions with lifelike avatars in 3D environments*.

Gulz, A., & Haake, M. (2006). Design of animated pedagogical agents- A look at their look. *International Journal of Human-Computer Studies, 64*, 322-339.

Harp, S. F., & Mayer, R. E. (1997). The role of interest in learning from scientific text and illustrations: On the distinction between emotional interest and cognitive interest. *Journal of Educational Psychology, 89*, 92–103.

Harp, S. F., & Mayer, R. E. (1998). How seductive details do their damage: A theory of cognitive interest in science learning. *Journal of Educational Psychology*, *90*, 414–434.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Holton, E.F. III. (1996). The flawed four-level evaluation model. *Human Resource Develoment Quarterly, 7*(1), 5-21.

Humphreys, J., Novicevic, M., Olson, W., & Ronald, M. (2010). Time to "hunker down?": Not necessarily. Drastic cuts in travel, benefits, and training can be devastating to long-term viability and success. http://www.businessweek.com/managing/content/jan2010/ca20100128_592971.htm.

Jansen, J.M., Daams, J.G., Koeter, M.W.J., Veltman, D.J., van den Brink, W., & Groudiaan, A.E. Effects of non-invasive neurostimulation on craving: A meta-analysis. *Neurosci Biobehav Review,* 2013.

Johnson, W. L., Rickel, J. W., & Lester, J. C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education, 11*(1), 47-78.

Johnson, W. L., Rickel, J. W., Stiles, R., & Munro, A. (1998). Integrating pedagogical agets into virtual environments. *Presence: Teleoperators and Virtual Environments, 7*(6), 523-546.

*Kim, Y. (2004). *Pedagogical agents as learning companions: The effects of agent affect and gender on learning, interest, self-efficacy, and agent persona* (Unpublished doctoral dissertation) Florida State University, Tallahassee, FL.

*Kim, Y. (2007). Desirable characteristics of learning companions. *International Journal of Artificial Intelligence in Education, 17*(4), 371-388.

*Kim, Y., Baylor, A. L., & PALS Group. (2006). Pedagogical agents as learning companions: The role of agent competency and type of interaction. *Educational Technology Research and Development, 54*(3), 223-243.

*Kim, Y., Baylor, A. L., & Shen, E. (2007). Pedagogical agents as learning companions: The impact of agent emotion and gender. *Journal of Computer Assisted Learning*, *23*(3), 220-234.

*Kim, C., Keller, J. M., & Baylor, A. L. (2007). Effects of motivational and volitional messages on attitudes toward engineering: Comparing text messages with animated messages delivered by a pedagogical agent. In *Proceedings of the IADIS International Conference of Cognition and Exploratory Learning in Digital Age (CELDA)* (pp. 317-320).

Kim, M. & Ryu, J. (2003). Meta-Analysis of the Effectiveness of Pedagogical Agent. In D. Lassner & C. McNaught (Eds.), Proceedings of ED-MEDIA 2003--World Conference on Educational Multimedia, Hypermedia & Telecommunications (pp. 479-486). Honolulu, Hawaii, USA: Association for the Advancement of Computing in Education (AACE). Retrieved September 11, 2017 from https://www.learntechlib.org/p/13806/.

Kirkpatrick, D.L. (1976). *Evaluation of training*. New York: McGraw-Hill.

*Ko, Y. A. (2010). *The effects of pedagogical agents on listening anxiety and listening comprehension in an English as a foreign language context.* Utah State University. All Graduate Theses and Dissertations. Paper 822.

Kraiger, K. (2002). *Decision-based evaluation*. San Francisco, CA: Jossey-Bass.

Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., & Bhogal, R. S. (1997). The persona effect: affective impact of animated pedagogical agents. In Proceedings of the ACM SIGCHI Conference on Human factors in computing systems (359-366). ACM.

Lester, J. C., Stone, B. A., & Stelling, G. D. (1999). Lifelike pedagogical agents for mixed-initiative problem solving in constructivist learning environments. *User Modeling and User-Adapted Interaction, 9,* 1-44.

Lester, J. C., Voerman, J. L., Towns, S. G., & Callaway, C. B. (1999). Diectic believability: Coordinated gesture, locomotion, and speech in lifelike pedagogical agents. *Applied Artificial Intelligence, 13,* 383-414.

Louwerse, M. M., Graesser, A. C., Lu, S., & Mitchell, H. H. (2005). Social cues in animated conversational agents. *Applied Cognitive Psychology, 19*, 693-704.

*Lusk, M. M., & Atkinson, R. K. (2007). Animated pedagogical agents: Does their degree of embodiment impact learning from static or animated worked examples?. *Applied Cognitive Psychology*, *21*(6), 747-764.

Mayer, R.E. (2001). *Multimedia learning*. New York: Cambridge University Press.

Mayer, R.E. (2003). The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and Instruction, 13*, 125-139.

Mayer, R.E. (2005). Cognitive theory of multimedia learning *The Cambridge handbook of multimedia learning* (pp. 31-48).

*Mayer, R. E., & DaPra, C. S. (2012). An embodiment effect in computer-based learning with animated pedagogical agents. *Journal of Experimental Psychology: Applied*, *18*(3), 239.

*Mayer, R. E., Dow, G. T., & Mayer, S. (2003). Multimedia learning in an interactive self-explaining environment: What works in the design of agent-based microworlds?. *Journal of Educational Psychology*, *95*(4), 806.

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology, 93*, 187–198.

Mayer, R.E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 43-52.

McCloud, S. (1993). *Understanding Comics*. New York, NY: HarperPerennial.

*Mitchell, T. D., Brown, S. G., Mann, K. E., & Towler, A. J. (2014). Three to tango: Agent, feedback comparison, and goal-orientation on training outcomes. Paper presented at the 29th annual conference of the Society of Industrial/Organizational Psychology Conference, Honolulu.

*Mitrovic, A., & Suraweera, P. (2000). Evaluating an animated pedagogical agent. In *Intelligent Tutoring Systems* (73-82). Springer Berlin/Heidelberg.

*Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents?. *Cognition and Instruction, 19*(2), 177-213.

Mori, M. (1970). Bukimi no tani. [The Uncanny Valley]. *Energy, 7*(4), 33-35.

*Moundridou, M., & Virvou, M. (2002). Evaluating the persona effect of an interface agent in a tutoring system. *Journal of Computer Assisted Learning*, *18*(3), 253-261.

Paivio, A. (1986). *Mental representations: A dual-coding approach*. New York: Oxford University Press.

*Park, S. (2005). *The effects of seductive augmentation and agent role on learning interest, achievement, and attitude*. Florida State University.

*Perez, R., & Solomon, H. (2005). Effect of a Socratic animated agent on student performance in a computer-simulated disassembly process. *Journal of Educational Multimedia and Hypermedia, 14*(1), 47.

Pigott, T.D., & Polanin, J.R. The use of meta-analytic statistical significance testing. *Research Synthesis Methods, 6*, 1:63-73, 2014. Retrieved from Loyola eCommons, School of Education: Faculty Publications and Other Works, http://dx.doi.org/10.1002/jrsm.1124

*Quesnell, T. J. (2012). Pedagogical training agents, supraliminal priming, and training outcomes. DePaul University. Chicago, IL.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places.* New York: Cambridge University Press.

*Rosenberg-Kima, R. B., Baylor, A. L., Plant, E. A., & Doerr, C. E. (2008). Interface agents as social models for female students: The effects of agent visual presence and appearance on female students' attitudes and beliefs. *Computers in Human Behavior*, *24*(6), 2741-2756.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*(3), 638-641.

Salas, E., & Cannon-Bowers, J.A. (2001). The science of training: A decade of progress. *Annual Review of Psychology, 52*, 471-499.

Schroeder, N. L., Adesope, O. O., & Gilbert, R. B. (2013). How effective are pedagogical agents for learning? A meta-analytic review. *Journal of Educational Computing Research*, *49*(1), 1-39.

Schulman, A. I. (1974). Memory for words recently classified. *Memory & Cognition, 2*(1A), 47-52.

Shaw, E., Johnson, W. L., & Ganeshan, R. (1999). *Pedagogical agents on the web*. Paper presented at the International Conference on Autonomous Agents, Seattle, WA. http://www.speakeasydesigns.com/SDSU/student/640/agents99-draft.pdf

*Shiban, Y., Schelhorn, I., Jobst, V., Hörnlein, A., Puppe, F., Pauli, P., & Mühlberger, A. (2015). The appearance effect: Influences of virtual agent features on performance and motivation. *Computers in Human Behavior, 49*, 5-11.

*Son, C. (2009). *The effects of pedagogical agent-delivered persuasive messages with fear appeal on learners' attitude change*. The Florida State University.

Sugrue, B., & Rivera, R.J. (2005). State of the industry: ASTD's annual review of trends in workplace learning and performance. Alexandria, VA: American Society for Training and Development.

Swanson, R.A., & Falkman, S.K. (1997). Training delivery problems and solutions: Identification of novice trainer problems and expert trainer solutions. *Human Resource Development Quarterly, 8*(4), 305-314.

Tannenbaum, S.I., Cannon-Bowers, J.A., Salas, E., & Mathieu, J.E. (1993). Factors that influence training effectiveness: A conceptual model and longitudinal analysis.

*Theodoridou, K. (2011). Learning Spanish with Laura: the effects of a pedagogical agent. *Educational Media International*, *48*(4), 335-351.

*Towler, A.J., Arman, G., Quesnell, T.J., & Hofmann, L. (2014). How charismatic trainers inspire others to learn through positive affectivity. *Computers in Human Behavior*, *32*, 221-228.

Towler, A.J., Kraiger, K., Sitzmann, T., Van Overberghe, C., Kuo, J., Ronen, E., & Stewart, D. (2008). The seductive details effect in technology-delivered instruction. *Performance Improvement Quarterly*, *21*(2), 65-86.

*Unal-Colak, F., & Ozan, O. (2012). The effects of animated agents on students' achievement and attitudes. *Turkish Online Journal of Distance Education*, *13*(2), 96-111.

*Veletsianos, G. (2009). The impact and implications of virtual character expressiveness on learning and agent–learner interactions. *Journal of Computer Assisted Learning*, *25*(4), 345-357.

*Veletsianos, G. (2010). Contextually relevant pedagogical agents: Visual appearance, stereotypes, and first impressions and their impact on learning. *Computers & Education, 55*, 576-585.

*Veletsianos, G. (2012). How do learners respond to pedagogical agents that deliver social-oriented non-task messages? Impact on student learning, perceptions, and experiences. *Computers in Human Behavior*, *28*(1), 275-283.

Veletsianos, G., & Miller, C. (2008). Conversing with pedagogical agents: A phenomenological exploration of interacting with digital entities. *British Journal of Educational Technology, 39*(6), 969-986.

*Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., & Collins, H. (2008). The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human-Computer Studies, 66*(2), 98-112.

Weizenbaum, J. (1966). ELIZA – a computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36-45.

Yildiz, A., Vieta, E., Leucht, S., & Baldessarini, R. J. (2011). Efficacy of antimanic treatments: Meta-analysis of randomized, controlled trials. *Neuropsychopharmacology*, *36*(2), 375.

*Yung, H. I., & Dwyer, F. M. (2010). Effects of an animated agent with instructional strategies in facilitating student achievement of educational objectives in multimedia learning. *International Journal of Instructional Media*, *37*(1), 55-65.

*Zumbach, J., Schmitt, S., Reimann, P., & Starkloff, P. (2006). Learning life sciences: Design and development of a virtual molecular biology learning lab. *The Journal of Computers in Mathematics and Science Teaching*, *25*(3), 281.

# Appendix A. Code Book

| | |
|---|---|
| **Coder Name:** | |
| **Authors:** | |
| **Publication Year:** | **Study Name (eg., TQ001):** |

| **Publication Type (choose one):** | | |
|---|---|---|
| **___Journal Article (1)** | **___ Book Chapter (2)** | **__ Conf Paper/Presentation (3)** |
| **___ Masters Thesis (4)** | **___ Doctoral Dissertation (5)** | **__ Unpublished Manuscript (6)** |

| **1. STUDY CONTEXT (choose one)** |
|---|
| **_____ Lab (1)** |
| **_____ Applied/Real World (2)** |

| **2. Did the study use a student sample? (choose one):** |
|---|
| **_____ Yes (1)** |
| **_____ No (2)** |

| **3. SAMPLE DEMOGRAPHICS:** |
|---|
| **_____ Percent Male** |
| **_____ Percent White/Caucasian** |
| **_____ Average Age** |

| **4. STUDY DESIGN (choose one):** |
|---|
| **_____ Between-subjects (1)** |
| **_____ Within-subjects (2)** |
| **_____ Mixed (3)** |

| **5. Was there incentive to perform well?** |
|---|
| **_____ Yes (1)** |
| **_____ No (2)** |
| **If Yes, please describe the incentive:** |

| **6. Is there any dependent data in the study? (choose one):** |
|---|
| **_____ Yes (1)** |
| **_____ No (2)** |
| **If Yes, please explain below:** |

| **7. Does the learner practice the task/skill being taught? (select all that apply):** |
|---|
| **_____ No (1)** |
| **_____ Before training (2)** |
| **_____ During training (3)** |
| **_____ After training (4)** |

**8. HUMAN versus NON-HUMAN (choose one):**

*Was the pedagogical agent used in the study humanlike or not?*

_____ **Humanlike (1)-** Reasonably resembles a human being in form, ability, dimensions/size, and movement

_____ **Non-Humanlike (2)**- Does not reasonably resemble a human being in form, ability, dimensions/size, and movement

**If (2), please explain why:**

_____ **Cannot Determine (3)-** (Grounds for the study to be excluded from the analysis)

**9. DEGREE OF ICONICITY (choose one):**

*How cartoonlike or realistic is the agent?*

_____ **High (Cartoon) (1)**- These agents are cartoon-like, exhibit unnatural coloration, movement, lack of detail, blurriness, and/or other features that make it apparent the agent was created and is not real

_____ **Moderate (2)**- These agents exhibit equal amounts of high and low iconicity

_____ **Low (Realistic) (3)**- These agents are photorealistic, video animated, have high levels of detail in their animation, and/or incorporate high levels of fine lines and shading.

_____ **Cannot Determine (4)** (Grounds for the study to be excluded from the analysis)

**10. AGENT ROLE (I = included, P = primary):**

*Agents are utilized to provide support for learning, by definition. Which "Agent Role" describes the level of learning support provided by the agent?*

_____ **Supplanting (1)-** The agent *performs most tasks* for learners to illustrate successful task completion

_____ **Demonstrating (2)**- The agent *performs a task and then* observes the *learner perform* the task (allowing learners to practice the task)

_____ **Modeling (3)-** The agent demonstrates a task, and *articulates the rationale and strategies* being used to execute the task

_____ **Coaching (4)-** The agent provides *hints and feedback when the learner has trouble* throughout the execution of a task

_____ **Testing (5)-** The agent *challenges the learner's knowledge* about elements of a task to facilitate learning (before, during, or after task execution/explanation)

_____ **Cannot Determine (6)**

**11. INSTRUCTION MODALITY (I = included, P = primary):**

*Agents are utilized to provide support, by definition. Which "Instruction Modality" describes the type/style of learning support provided by the agent?*

_____ **Executing (1)**-The agent performs actions (e.g., task practice, activities, exercises, etc.) *instead of the learner* performing them

_____ **Showing (2)**- The agent provides demonstrations for the learner, later allowing the *learner to replicate* actions (in practice, not testing)

_____ **Explaining (3)**- The agent provides *feedback or clarifications* about a task *while learners perform* them (in practice, not testing)

_____ **Questioning (4)**- The *agent asks questions* about the task or elements of the task for learners to answer (at any point)

_____ **Cannot Determine (5)**

**12. SUPPORT ADAPTATION**
**(I = included, P = primary):**

*Support Adaptation refers to any actions taken by agents to help learners engage in effective processing of information during the training session. Agents can vary with regard to the Quantity of support they provide, and the Object on which their support focuses*

**Quantity**- The amount of support the agent provides a learner

_____ **No support (1)**- Only a learning goal and content are delivered (learners determine how to process information on their own)

_____ **Activation (2)**- Goal statements and content are delivered, and *a few tips/indicators* are provided regarding how to process the content (activating/stimulating learners' cognitive processing skills)

_____ **Compensation (3)**- Decisions related to cognitive processing are taken over from the learner; *frequent/multiple tips/indicators* are provided regarding how to process the content (the learner requires little thought/reflection to determine how to learn/process the content presented)

_____ **Cannot Determine (4)**

**Object**- Where the agent directs learner attention while supporting learning

_____ **Content (1)**- Teaching specific elements of the subject matter/topic (e.g., increasing knowledge)

_____ **Problem Solving (2)**- Teaching strategies used to solve a problem or complete a task (e.g., increasing skills/abilities needed to achieve a performance goal)

_____ **Meta-Cognition (3)**- Teaching learners to think about their thinking (e.g., highlighting course learning goals, self-monitoring, progress tracking, etc.)

_____ **Technology (4)**- Support related to the technology or tools used to complete a task

_____ **Cannot Determine (5)**

**13. TIMING OF SUPPORT DELIVERY**
**(I = included, P = Primary):**

_____ **Prior to Need (1)**- Presents information relevant task support/information before task execution

_____ **Just-In-Time (2)**- Presents relevant task support/information in a timely manner during task execution

_____ **Delayed (3)**- Provides example information about the task after task/practice completion

_____ **Cannot Determine (4)**

**14. CONTROL**
**(choose primary):**

_____ **Agent (1)**

_____ **Learner (2)**

| **15. DELIVERY MODALITY (P = Primary, choose one from each row):** |
|---|
| **_____ Speech (1) or _____ Text (2)** |
| **_____ Monologue (3) or _____ Personalized Messages (4)** |
| **DELIVERY MODALITY (I = included for all that apply):** |
| **_____ Facial Expressions (5)** |
| **_____ Gestures (6)** |

| **DEPENDENT VARIABLES (Use one sheet for each dependent variable measured)** |
|---|
| **What Criterion was used?** |
| **_____ Post-training Self-Efficacy (1)**- Self-efficacy/confidence measures implemented after the training |
| **_____ Cognitive Learning (2)**- Multiple choice, written, matching, or other test given to test |
| **_____ Training Performance (3)**- Participants are asked to practice the tasks/skills they just learned on a sample problem during or immediately after the training |
| **_____ Transfer Performance (4)**- Participants are asked to complete a task or use a skill taught in the training after an extended amount of time or in a real-world, applied setting |
| **_____ Other (5)**- Indicates the article includes other potentially relevant DVs (not 1-4) |

## DEPENDENT VARIABLE CODING SHEET

| |
|---|
| **Study Number (e.g., TQ001):** |
| **Criterion:** |
| **_____ Post-training Self-Efficacy (1)** |
| **_____ Cognitive Learning (2)** |
| **_____ Training Performance (3)** |
| **_____ Transfer Performance (4)** |

**Please indicate which type of data is being reported in the article:**
**_____Between Subjects Design -or- _____Within Subjects Design**

**_____N Control** **_____N Test**

**_____Mean Control** **_____Mean Test**

**_____SD Control** **_____SD Test**

**_____Effect Size(d) Control** **_____Effect Size(d) Test**

**_____*F* Stat Control** **_____*F* Stat Test**

**_____*t* Stat Control** **_____*t* Stat Test**

**_____Other**

| **Correlation Matrix: List Second Variable and Relevant Statistics Below** | | | | | |
|---|---|---|---|---|---|
| **Criterion** | **N** | **Control** | **Experimental** | **z (diff in r's)** | **d (effect size)** |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |

## Appendix B. Article List

| Article List and Publication Type | | |
|---|---|---|
| **Count** | **Article** | **Publication Type** |
| 1 | Atkinson (2002) | Journal Article |
| 2 | Baylor & Kim (2004) | Unpublished Manuscript |
| 3 | Baylor & Kim (2005) | Journal Article |
| 4 | Baylor & Kim (2005) | Journal Article |
| 5 | Baylor & Ryu (2003) | Journal Article |
| 6 | Baylor, Kim, & Shen (2010) | Journal Article |
| 7 | Choi & Clark (2006) | Journal Article |
| 8 | Clarebout & Elen (2005) | Journal Article |
| 9 | Craig & Gholson (2002) | Journal Article |
| 10 | Craig, Driscoll, & Gholson (2004) | Journal Article |
| 11 | Dirkin, Mishra, & Altermatt (2005) | Journal Article |
| 12 | Domagk (2010) | Journal Article |
| 13 | Dunsworth & Atkinson (2007) | Journal Article |
| 14 | Fujimoto (2010) | Doctoral Dissertation |
| 15 | Kim & Baylor (2006) | Journal Article |
| 16 | Kim (2004) | Doctoral Dissertation |
| 17 | Kim (2007) | Journal Article |
| 18 | Kim, Keller, & Baylor (2007) | Conference Presentation |
| 19 | Ko (2010) | Doctoral Dissertation |
| 20 | Liew, Tan, & Jayothisa (2013) | Journal Article |
| 21 | Lusk & Atkinson (2007) | Journal Article |
| 22 | Mayer & DaPra (2012) | Journal Article |
| 23 | Mayer, Dow, & Mayer (2003) | Journal Article |
| 24 | Mitchell. Brown, Mann, & Towler (2014) | Journal Article |
| 25 | Mitrovic & Suraweera (2000) | Journal Article |
| 26 | Moundridou & Virvou (2002) | Journal Article |
| 27 | Park (2005) | Doctoral Dissertation |
| 28 | Perez & Solomon (2005) | Journal Article |
| 29 | Quesnell (2012) | Unpublished Manuscript |

| | | |
|---|---|---|
| 30 | Rosenberg-Kima, Baylor, Plant, Doerr (2008) | Journal Article |
| 31 | Shiban, Schelhorn, Jobst, Hornlein, Puppe, Pauli, Muhlberger (2015) | Journal Article |
| 32 | Son (2009) | Doctoral Dissertation |
| 33 | Theodoridou (2011) | Journal Article |
| 34 | Towler, Arman, Quesnell, Hoffman (2014) | Journal Article |
| 35 | Unal-Colak & Ozan (2012) | Journal Article |
| 36 | Veletsianos (2009) | Journal Article |
| 37 | Veletsianos (2010) | Journal Article |
| 38 | Veletsianos (2012) | Journal Article |
| 39 | Wang, Johnson, Mayer, Rizzo, Shaw, & Collins (2008) | Journal Article |
| 40 | Yung & Dwyer (2010) | Journal Article |
| 41 | Zumbach, Schmitt, Reimann, & Starkloff (2006) | Journal Article |