# A Meta-Analysis of the Impact of the Inclusion and Realism of Human-Like Faces on User Experiences in Interfaces

**Nick Yee, Jeremy N. Bailenson, Kathryn Rickertsen**
Department of Communication
Stanford University, Stanford, CA
{nyee, bailenson, kathrynr}@stanford.edu

**ABSTRACT**

The use of embodied agents, defined as visual human-like representations accompanying a computer interface, is becoming prevalent in applications ranging from educational software to advertisements. In the current work, we assimilate previous empirical studies which compare interfaces with visually embodied agents to interfaces without agents, both using an informal, descriptive technique based on experimental results (46 studies) as well as a formal statistical meta-analysis (25 studies). Results revealed significantly larger effect sizes when analyzing subjective responses (i.e., questionnaire ratings, interviews) than when analyzing behavioral responses such as task performance and memory. Furthermore, the effects of adding an agent to an interface are larger than the effects of animating an agent to behave more realistically. However, the overall effect sizes were quite small (e.g., across studies, adding a face to an interface only explains approximately 2.5% of the variance in results). We discuss the implications for both designers building interfaces as well as social scientists designing experiments to evaluate those interfaces.

**Author Keywords**
Computer-mediated communication, quantitative methods, meta-analysis, embodied agents, realism.

**ACM Classification Keywords**
H5.m. Information interfaces and presentation (e.g., HCI): H.5.2 User Interfaces.

## INTRODUCTION

*Embodied agents* are visual digital representations of a computer interface often in the form of human-like faces [13]. They typically either accompany or replace traditional computer interfaces and are becoming more and more prevalent in military applications [52], video games [14], online learning systems [33], advertisements [31] and even on cellular phones [54]. Over the past decade, much empirical research has been dedicated towards examining the extent to which these embodied agents improve an interaction with an interface, beginning with an early landmark paper by Walker, Sproull, and Subramani [59]. Although many researchers have examined the presence and the type of embodied agents, there is little consensus as to whether or not the presence of visual agents improves a user's experience with an interface, and if so by what degree.

An early attempt to provide an overview of the literature was offered by Dehn and van Mulken [17], who descriptively examined eight independent studies in terms of the variables manipulated and the outcome measures. They reported mixed results: adding an embodied agent to an interface made the experience more entertaining according to survey measures, but made no difference in the ratings of the quality of the agent. In terms of behavioral responses, the few studies available at the time showed no consistent differences in a user's performance between the two types of interfaces.

The goal of the current paper is to use both informal (i.e., similar to Dehn and van Mulken [17]) as well as formal (i.e., a statistical meta-analysis) techniques to summarize the previous work which adds embodied agents to human-computer interfaces. Scholars examining human-computer interaction have often used meta-analyses to assimilate the research. Previous papers presented at CHI have used both informal, descriptive techniques to explain the literature (e.g., assimilating design approaches [42]) and more formal statistical techniques as well [63]. This latter paper examined 39 studies that compared interviews either administered in person or via computer, and demonstrated that more personal disclosure occurred in front of a computer than in front of a live person. Moreover, by examining the effects over time, their analyses traced the design changes in interfaces and provided practical implications for building new applications.

Using the meta-analysis as a way to synthesize research in an area, we examined a number of theoretical questions

related to interfaces: the effect of the presence of an embodied agent, how realistic the agent appears through animations and behaviors, and the type of response which was measured (subjective or performance). The current work seeks to provide a thorough review of the literature which has examined embodied agents as interface agents and to provide mathematical summaries of the extent to which embodied agents improve interfaces under various circumstances.

There is much controversy about what constitutes realism in an interface (see [9] for an in-depth discussion of these ideas). Clearly there are many dimensions on which an interface agent can be considered real—it can behave realistically through animations, it can be highly photographically realistic via computer graphics, or alternatively it can be highly humanlike (i.e., anthropomorphic). In our meta-analysis, we were primarily interested in the presence of visual human representations, and the realism level of those representations. In the current work, we excluded a small number of studies that specifically examined non-human interface agents such as animals (see [40] for a discussion of human-nonhuman interfaces). Consequently, the current operationalization of realism is specifically defined as being more realistic on either the behavioral or photographic dimensions of realism (or both). Unfortunately, we did not have a high enough number of papers to conduct systematic comparisons of these two types of realism, because researchers most often manipulated both simultaneously.

Our general research interest could thus be elaborated as the following: Do people react differently to interfaces with 1) no visual representation, 2) a human-like representation with low realism (e.g., cartoon figure), and 3) a human-like representation with high realism (e.g., 3D model animated with gestures)? Thus, we were primarily interested in the concept of realism rather than anthropomorphism (i.e., the degree to which something resembles a human).

There are reasons to assume that adding an embodied agent will improve an interaction with a user. For example, Takeuchi and Naito [53] point out the embodied agent may draw the attention of the user and make him or her more engaged. However, the danger of this addition is that it may distract the user from the very task on which the agent is supposedly aiding. Clearly the design goal when adding an interface agent is to make the visual representation improve task performance, but there is no clear blueprint for creating embodied agents that keep users focused on the content of the interface. In the current work we examined a large number of studies to determine whether, overall, embodied agents tend to augment, distract, or have no effect on task performance.

### Hypotheses

The descriptive results reported by Dehn and van Mulken [17], as well as the overall reporting of the field, suggested several observed effects that we could examine more precisely. First of all, the inclusion of any visual representation seems to improve task performance when compared with not having a visual representation at all. And secondly, animated agents with higher realism seem to lead to higher task performance than agents with lower realism.

Research in this area has primarily employed two kinds of task performance measures - subjective measures and behavioral measures. Over the past decade, some researchers have criticized the validity of subjective measures, particularly questionnaires that are meant to measure the effectiveness of embodied agents and virtual humans. For example, Slater [50] has shown that even meaningless questions can produce seemingly valid and reliable results when used to describe a virtual experience. In other studies [5], large differences in behavioral measures, such as the amount of mutual gaze a person maintained with a virtual representation of another, were not reflected in self-report surveys of social presence. And finally, subjective measures may provide an accessible outlet for what psychologists term *demand characteristics*, the phenomenon of experimental subjects conforming to or obstructing the study hypotheses, especially when research objectives become apparent to participants. In other words, there is reason to believe that subjective measures and performance measures may produce very different results, even in the same study. Thus, we could restate the above-mentioned observed effects as more specific hypotheses:

H1a. The inclusion of any visual representation will lead to higher behavioral task performance measures than having no visual representation.

H1b. The inclusion of any visual representation will lead to higher subjective task performance measures than having no visual representation.

H2a. Animated agents with higher realism will lead to higher behavioral task performance measures than agents with lower realism.

H2b. Animated agents with higher realism will lead to higher subjective task performance measures than agents with lower realism.

### Research Questions

The availability of both subjective and behavioral measures also allows us to explore whether subjective measures produced different results than behavioral measures, however, we did not have specific predictions for how the two may differ. Thus, one research question was:

RQ1. Given the observed mismatches between subjective and behavioral measures in the literature, are there overall differences in effect sizes between these two kinds of dependent variables?

And finally, technology has changed rapidly over the past decade and social norms may differ between different cultures as new technologies are introduced. Moreover,

some of the studies were conducted in immersive virtual reality while others were conducted using desktop equipment. Thus, we were also interested in whether study results were significantly influenced by their year of publication, location of study, or the equipment used.

RQ2. Does year of publication have any impact on the findings of the study?

RQ3. Does the location where the research was conducted have any impact on the findings of the study?

RQ4. Do studies conducted in immersive virtual reality produce different results from those conducted in a desktop setting?

## METHOD

### Selection of Studies

The studies considered for inclusion in this analysis were culled from bibliographic indexes related to the fields of psychology, computer-mediated communication (CMC), and virtual reality. These included Expanded Academic ASAP, Google Scholar, Google keyword, PsycInfo, PsycArticles Fulltext Search, InterDok, ProQuest, and SearchPlus. In this initial pass, articles that appeared to report an experimental study of anthropomorphism, embodied agents, or agent realism were collected and reviewed. Sources were only considered if they were published in a peer-reviewed journal or in published conference proceedings. This ensured a basic level of methodological and data integrity in the pool of included studies. On the other hand, this potentially leads to a bias towards studies that showed results that were significantly different from the null hypothesis. We will return to this issue again in the discussion section.

The literature review yielded 106 studies. Several selection criteria were then applied. First, an article was included only if it was an experimental study that manipulated the variables of interest and contained clear reports of quantitative data relating to the outcome of different conditions. Thus, purely qualitative studies involving open-ended self-reports or observational user studies without quantitative coding schemes or dependent variables were removed.

In many cases, articles described experimental studies involving dependent variables, but did not report the statistics needed for the formal meta-analysis. For example, a study might report the ANOVA F-value for the outcome of three conditions without reporting the means and standard deviations of the individual conditions. In these cases, it would not be possible to generate an effect size value if we needed to compare one of the three conditions against another. We discuss the details of the necessary statistics in more detail in the next section. For each experimental study that clearly measured dependent variables but did not report specific statistics in the article, we individually contacted the lead authors of the study via email in an effort to gather those statistics.

Of the 106 articles, 61 were discarded because they were outside the scope of the study (e.g., a manipulation of voice with no visual components, or compared agents embodied as animals to ones embodied as humans), were theoretical articles with no empirical data, or were qualitative studies without quantitative measures. Of the 46 papers remaining, 25 provided enough data to be included in the formal meta-analysis (or we were able to get enough information via personal correspondence with the authors), while the other 17 did not provide enough statistics (usually standard deviations) to be included. Although this appears to be a low number of studies, previous meta-analytic studies in computer-mediated communication have also tended to be based on only about a dozen useable studies (for example, see [61]).

For each study, we coded: 1) means and standard deviations of all relevant conditions according to the mentioned comparison conditions, 2) number of participants in each condition, 3) year of publication, 4) country where the study was conducted (if not available, we used the country of the affiliated institution of the primary author), and 5) platform on which the study was conducted (i.e., desktop or immersive).

Of these 25 studies, the average year of publication was 2001.96 (SD = 2.29) with a median of 2002. The average sample size within each study was 45.40 (SD = 35.55). With regard to study location, 13 were conducted in the US or Canada, 9 were performed in Europe, and the remaining 3 were conducted in Asia. And finally, with regard to equipment used, 17 were conducted on desktop equipment, 6 were conducted using immersive virtual reality, and the remaining 2 were conducted on a large projected screen.

### Effect Size Calculations

To generate the necessary effect size tabulations in order to test our hypotheses, we tabulated several possible effect sizes for each paper depending on the available conditions. First, we tabulated the results of performance data separately from the results of subjective data. Performance data might include time to task completion, accuracy measures, or similar behavioral measures. Subjective data, on the other hand, was any measure that was based on self-report or survey data. Second, we tabulated effect sizes based on two kinds of comparisons between conditions. We wanted to be able to look at the effect of no representation against any degree of representation independently from the effect of low against high realism. In other words, for each included study, there were four potential effect sizes that could be calculated depending on the available conditions and dependent variables - 1) subjective measures of representation vs. no representation, 2) behavioral measures of representation vs. no representation, 3) subjective measures of high vs. low realism, and 4) behavioral measures of high vs. low realism. We refer to these as the

*four comparison conditions* in the remainder of the paper. Also note that for studies which reported more than one experiment, effect sizes were calculated for each experiment separately and each experiment counted as its own case for the meta-analysis.

This tabulation process might be clearer with an example. In a hypothetical experimental study, participants are assigned to interact with an agent with either 1) just text, 2) with a cartoon face, or 3) with a highly photorealistic face. During the task, participants are timed for performance. After the task, they are asked to judge the friendliness and appeal of the agent on a survey. In this example, to generate the effect size of the performance differential of no representation against any degree of representation, we would collapse the cartoon and photorealistic face condition and compare the averaged score of the performance times against the score of the textual condition. The other three possible tabulations are derived accordingly. Of course, in many studies, the available conditions do not allow the generation of the effect sizes from all four comparison conditions. For each study, we computed as many of the four comparison conditions as was possible.

We calculated these effect sizes based on formulas described in the widely referenced work by Rosenthal [47]. The effect size variable *r* is a measure of the impact that a manipulation has on a dependent variable. Squaring this variable to get $r^2$ shows the amount of variance in the dependent measure that can be accounted for by the manipulation. For example, an $r^2$ of .15 means that 15% of the variance in the dependent measure can be explained by a manipulation in realism (or whatever the relevant manipulation is). For each study, we calculated an effect size *r* value for each possible comparison related to our hypotheses. In a study with only two relevant conditions (i.e., no representation against some representation), it is possible to derive the *r* value from a *t* value along with the degrees of freedom. If the *t* value is not reported, the availability of the means and standard deviations along with the number of cases in each condition would allow the derivation of the *t* value, and thus the *r* value. In a study with multiple conditions, it is not possible to generate a relevant *r* value from an ANOVA *F* value because the omnibus *F* value doesn't test the specific comparisons of interest to us. Thus, in these cases, we used the means and standard deviations, if reported, to generate the *r* value via a *t* value.

In many cases, a study may contain multiple dependent variables of interest. For example, a study may use a variety of performance measures. In these cases, we first calculated the *r* values independently for each measure. As described in Rosenthal [47], the averaging of *r* values must be performed via a *z* transformation because the *r* distribution does not follow a normal distribution. Thus, the individual r values are transformed into *z* values, averaged, and then the averaged *z* value is transformed back into the new aggregate *r* value.

The sign of each *r* value describes the positive or negative effect of the comparison and this was kept constant throughout the meta-analysis. Thus, a positive *r* value signifies a positive increase in performance or subjective rating when comparing no representation against some form of representation, or low against high realism, while a negative *r* value signifies a decrease.

Finally, once all the effect sizes for each study had been calculated, we calculated an overall effect size for each of the four comparison conditions by converting the *r* values into *z* values. We then averaged the *z* values after weighting them by the sample size of the study. The averaged *z* value was then converted back into the overall *r* value.

### Significance Value Calculations

The significance of an effect size is independent of the final *r* value. For example, a collection of large sample studies may yield a highly significant, but small *r* value. To tabulate the overall significance value, we converted the *t* values of each relevant comparison to a *z* value because the *z* distribution is normal. We then calculated a significance *z* value for each effect size calculated as described in the above section (as described in [36]). In studies where multiple measures were used, we derived each *z* value from each *t* value individually before averaging the overall *z* value for that particular comparison. For the aggregate significance level for each of the four comparison conditions, we used the Stouffer method [36], summing the z values for a given comparison condition (after weighting the values by the corresponding sample size), and then dividing by the square-root of the number of studies in that condition.

## RESULTS

### Formal Meta-Analyses

The results of the effect size and significance value aggregation are listed in Appendix A for each individual study and the overall values. The overall effect sizes of the four comparison conditions ranged from -.04 to .14. While three of the four comparison conditions were highly significant at *p* levels of less than .05, the comparison of high-low realism using performance measures was not significant, with $p = .14$.

We were also interested in whether the effect sizes in the studies varied as a function of other factors, such as whether a subjective or behavioral measure was used. To this end, we carried out a series of contrasts on the set of effect sizes based on other factors. In cases where a study had effect sizes from several comparison conditions, the effect sizes were first averaged according to the factors of interest. Thus, each study only contributed at most once to each factor.

First, we compared the effect sizes based on subjective measures against those based on performance measures. Studies using subjective measures had significantly larger effect sizes ($n = 26$, $r = .16$) than studies using performance

measures ($n$ = 17, $r$ = .09), $z$ = 2.37, $p$ = .02. In other words, participants indicated larger differences via subjective self-report than were observed via performance measures.

In our meta-analysis, we had also separated: 1) studies that compared interacting with an agent that had no facial representation versus an agent that had a facial representation (i.e., the yes-no comparisons), and 2) studies that compared interacting with faces of low realism versus faces of high realism (i.e., the high-low comparison). A comparison of these two groups of effect sizes revealed that the effect sizes from yes-no comparisons ($n$ = 25, $r$ = .16) were significantly larger than those from the high-low comparison ($n$ = 18, $r$ = .07), $z$ = 2.43, $p$ = .02.

We also compared the effect sizes of studies that were conducted on a desktop computer against those that were conducted in an immersive virtual reality environment. The effect sizes of studies conducted in immersive virtual reality ($n$ = 4, $r$ = .32) was approaching significance on being larger than those conducted on a desktop computer ($n$ = 25, $r$ = .12), $z$ = 1.85, $p$ = .06.

We also compared studies published before the year 2000 and those afterwards as a crude measure of whether technological advances have impacted the results of studies in the area. The contrast was not significant, $z$ = 1.36, $p$ = .17. And finally, we conducted a series of contrasts to compare the effect sizes of studies conducted in North America, Europe, and Asia. Again, we did not find a significant difference, $p$'s > .10.

### Informal Analyses

To provide a more thorough, albeit less rigorous, aggregation of the studies we found during the literature review, we decided to revisit the original data in the set of 46 selected papers and tabulate trends descriptively even if they did not provide the values necessary to make statistical comparisons. The lack of two kinds of data prevented us from making the statistical comparisons for the formal meta-analysis. For example, let us assume a data set that provided four pairs of dependent measures in one of our four comparison conditions. In many cases, these four measures might vary in scale and also standard deviation. Unless the paper reported standard deviations for those measures, it would be impossible to accurately aggregate them. In some cases, papers did not even report the underlying range of scale points a measure was based on.

To work around these limitations, we devised a crude comparison technique. For each pair of dependent measures in each of our comparison conditions, we simply counted the number of times each condition (i.e., low realism versus high realism) had the better score. We would then label the result with whichever condition had more tallies. So in our example, consider the example in which we were working with the comparison of low versus high realism for subjective measures. We might find that in three of the four dependent measures, the high realism condition scored more positively, while the low realism condition only did so once. In this case, we would indicate that the high realism condition produced more positive results. In Appendix B, we list these descriptive results from the available studies. These descriptive results are striking in that they predominantly show that interface agents have positive results on users across all the comparison conditions - implying more consistent and stronger effect sizes than we found in the formal meta-analysis. We will discuss this disparity in greater detail in the next section.

## DISCUSSION

The main comparison conditions in the formal meta-analysis produced several consistent findings. First, the presence of a representation produced more positive social interactions than not having a representation. This effect was found in studies that used both subjective and behavioral measures. Secondly, human-like representations with higher realism produced more positive social interactions than representations with lower realism; however, this effect was only found when subjective measures were used. Behavioral measures did not reveal a significant difference between representations of low and high realism.

In addition, we found several interesting differences via contrasts between studies with different features. For example, effect sizes tended to be larger when subjective measures were used than when behavioral measures were used. There are several potential explanations for this. It may be because subjective measures are more sensitive than performance measures. For example, task performance may be a less sensitive measure of attitudes towards an agent than a direct survey item. However, the opposite has also been shown. Differences in behaviors are sometimes undetected by direct survey items [5, 39], thus the difference we found may also be driven by demand characteristics. Participants interacting with an animated character (as opposed to a photograph) may suppose that the researcher is expecting a high appraisal. Unfortunately, the data from our meta-analysis is unable to tease out these potential explanations.

It was also interesting that the effect sizes in the yes-no comparisons were larger than the effect sizes in the low-high comparisons. This difference suggests that while the presence of a face is better than no face at all, the quality of the face matters much less. One limitation to interpreting this difference (and in fact, a limitation to research in this area in general), however, is the potential of confounding variables in digital renditions of human faces. In other words, it is quite possible that animating highly realistic faces inherently allows for residual attributes of the faces that are negative—for example making 3D human faces may produce gestures and animations that appear unnatural or disturbing [e.g., the uncanny valley effect, see 35]. In other words, it is not clear that a 3D face differs from a photograph or a 2D cartoon simply on the dimension of

realism or anthropomorphism. Thus, the difference we found may be magnified if our analyses could reflect this unexplored potential confound of further defining anthropomorphism or realism in this area of research.

Comparing the results from the formal meta-analysis to the descriptive informal table is illuminating. When assimilating the general findings, more than three-quarters of the published papers in our sample indicated that interface agents have positive effects on users. Given such an overwhelming majority, one might expect this to be an extremely large and consistent effect. However, when conducting the formal meta-analysis, we see that the manipulation with the largest effects (i.e., the subjective report data) accounts for less than three percent of all the variance across the studies. Consequently, one must be cautious when generalizing from a large number of published studies without taking into account the effect size. In other words, while most studies have found that interface agents have positive effects on task performance, these effects are overall actually quite small.

While we gave our best effort at being exhaustive in our literature review, it is of course highly likely that we failed to find and include other relevant studies in this area. On the other hand, the studies in our meta-analysis are probably representative of the large majority of studies in this area. A common critique of meta-analyses is the file drawer problem. In fields where significant differences from hypothesized nulls are favored by journals, it is assumed that a great deal of null results go unpublished. The fail-safe *n* is a measure of the number of such non-significant studies that would be needed to make our finding non-significant. In our case, when examining the ratio of studies which found significant results, we computed the lowest fail-safe *n* to be 100 in any of the four comparison conditions. Thus, this implies that our findings are likely to be stable given that the fail-safe *n* is about four times the number of published studies we were able to find.

Another potential concern with meta-analyses is that they combine studies with widely varying tasks and dependent measures, and thus it is not clear what aggregating them actually means. Rosenthal and DiMatteo [48] have addressed this "apples and oranges" critique by noting that "it can be argued, however, that it is a good thing to mix apples and oranges, particularly if one wants to generalize about fruit, and that studies that are exactly the same in all respects are actually limited in generalizability" (pg. 68). Also note that the studies included are actually moderately similar – all of them ask participants to interact with an agent while some measure of task performance is being tracked. Rosenthal and DiMatteo also note that "when studies vary methodologically, well-done meta-analyses take these differences into account by treating them as moderator variables" (pg. 68). In our meta-analysis, we were careful to examine methodological differences (e.g., behavioral vs. subjective measures) that we felt might impact results across studies.

There are implications for designers and researchers that derive from our meta-analysis. For designers, the meta-analysis makes it clear that a visual representation of an agent leads to more positive social interaction than not having a visual representation. On the other hand, it appears that the realism of the embodied agent may matter very little. For researchers, the differences between subjective and performance measures highlights the danger of interpreting results from only one type of measure. For example, it is surprising that subjective measures of the high-low realism conditions show a highly significant effect while the performance measures show no effect at all. As we've mentioned before, this may be due to the lower sensitivity of performance measures or due to demand characteristics in lab experiments. Future studies should be careful to include both types of measures to further our understanding of this mismatch.

## AUTHOR NOTES

The authors would like to thank Julie Farrell, Alice Siu, and Zi Lin for their help in acquiring and coding relevant papers. The authors would also like to thank Cliff Nass for his guidance on this project and Henriette Van Vugt for providing helpful comments on an earlier draft of this paper. The current work was partially supported by NSF grant 0527377. Detailed descriptions of how the individual studies were coded can be found at http://vhil.stanford.edu.

## REFERENCES

1. Bailenson, J., Blasovich, J., Beall, A., and Loomis, J., Equilibrium Theory Revisited: Mutual Gaze and Personal Space in Virtual Environments. *Presence*, 2001. 10(6): p. 583-598.
2. Bailenson, J., Beall, A., and Blasovich, J., Gaze and Task Performance in Shared Virtual Environments. *The Journal of Visualization and Computer Animation*, 2002. 13: p. 313-320.
3. Bailenson, J., Blasovich, J., Beall, A., and Loomis, J., Interpersonal Distance in Immersive Virtual Environments. *Personality and Social Psychology Bulletin*, 2003. 29: p. 1-15.
4. Bailenson, J., Swinth, K., Hoyt, C., Persky, S., Dimov, A., and Blascovich, J., The independent and interactive effects of embodied agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in Immersive Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 2005. 14(379-393).
5. Bailenson, J. and Yee, N., A Longitudinal Study of Task Performance, Head Movements, Subjective Report, Simulator Sickness, and Transformed Social Interaction in Collaborative Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 2006. 15.

6. Bartneck, C., How Convincing is Mr. Data's Smile: Affective Expressions of Machines. *User Modeling and User-Adapted Interaction*, 2001. 11: p. 279-295.
7. Bente, G., Kramer, N., Petersen, A., and Ruiter, J., Computer Animated Movement and Person Perception: Methodological Advances in Nonverbal Behavior Research. *Journal of Nonverbal Behavior*, 2001. 25(3): p. 151-166.
8. Beun, R-J., de Vos, E., and Witterman, C., Embodied conversational agents: Effects on memory performance and anthropomorphisation, in *IVA 2003, Lecture Notes in Computer Science 2792*, T. Rist, Editor. 2003. p. 315-319.
9. Blascovich, J., Loomis, J., Beall, A., Swinth, K., Hoyt, C., and Bailenson, J., Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry*, 2002. 13: p. 103-124.
10. Bonito, J., Burgoon, J., and Bengtsson, B. The Role of Expectations in Human-Computer Interaction. in *GROUP '99: International Conference on Supporting Group Work*. 1999. Phoenix, AZ.
11. Burgoon, J., Bengtsson, B., Bonito, J., Ramirez, A.J., and Dunbar, N. Designing Interfaces to Maximize the Quality of Collaborative Work. in *Hawaii International Conference on System Sciences*. 1999. Maui, Hawaii.
12. Burgoon, J., Bonito, J., Bengtsson, B., Cederberg, C., Lundeberg, M., and Allspach, L., Interactivity in Human-Computer Interaction: A Study of Credibility, Understanding, and Influence. *Computers in Human Behavior*, 2000. 16: p. 553-574.
13. Cassell, J., Nudge Nudge Wink Wink: Elements of Face-to-Face Conversation for Embodied Conversational Agents, in *Embodied Conversational Agents*, J. Cassell, Editor. 2000, MIT Press: Cambridge, MA. p. 1-27.
14. Cassell, J., Towards a Model of Technology and Literacy Development: Story Listening Systems. *Journal of Applied Developmental Psychology*, 2004. 25: p. 75-105.
15. Chan, F.Y. and Khalid, H.M., Is Talking to an Automated Teller Machine Natural and Fun? *Egronomics*, 2003. 46: p. 1386-1407.
16. Cowell, A.J., *Increasing the Credibility of Anthropomorphic Computer Characters: The Effects of Manipulating Nonverbal Interaction Style and Demographic Emodiment*. 2001, University of Central Florida: Orlando, Florida.
17. Dehn, D. and van Mulken, S., The impact of animated interface agents: a review of empirical research. *International Journal of Human Computer Studies*, 2000. 52: p. 1-22.
18. Fabri, M., Moore, D., and Hobbs, D. Expressive Agents: Non-Verbal Communication in Collaborative Virtual Environments. in *The Autonomous Agents and Multi-Agent Systems*. 2002. Bologna, Italy.
19. Gerhard, M., Moore, D., and Hobbs, D., Close Encounters of the Virtual Kind: Agents Simulating Copresence. *Applied Artifical Intelligence*, 2005. 19: p. 393-412.
20. Guadagno, R., Blascovich, J., Bailenson, J., and McCall, C., Virtual Humans and Persuasion: The Effects of Agency and Behavioral Realism. *Media Psychology*, in press.
21. Gulz, A., Social Enrichment by Virtual Characters - Differential Benefits. *Journal of Computer Assisted Learning*, 2005. 21: p. 405-418.
22. Hess, T.J., Fuller, M.A., and Mathew, J., Involvement and Decision-Making Performance with a Decision Aid: The Influence of Social Multimedia, Gender, and Playfulness. *Journal of Management Information Systems*, 2006. 22(3): p. 15-54.
23. Hongpaisanwiwat, C. and Lewis, M. Attentional Effect of Animated Character. in *Human-Computer Interaction - INTERACT*. 2003. Zurich, Switzerland: IOS Press.
24. Hook, K., Persson, P., and Sjolinder, M., Evaluating Users' Experience of a Character-Enhanced Information Space. *AI Communications*, 2000. 13: p. 195-212.
25. Kiesler, S., Waters, K., and Sproull, L., A Prisoner's Dilemma Experiment on Cooperation with People and Human-Like Computers. *Journal of Personality and Social Psychology*, 1996. 70: p. 47-65.
26. Koda, T. and Maes, P. Agents with Faces: The Effect of Personification of Agents. in *Human-Computer Interaction*. 1996. London, UK.
27. Koda, T., User Reactions to Anthropmorphized Interfaces. *IEICE Transactions on Information and Systems*, 2003. E86-D: p. 1369-1377.
28. Kramer, N. Social Communicative Effects of a Virtual Program Guide. in *Intelligent Virtual Agents*. 2005. Kos, Greece: Springer-Verlag Berlin Heidelberg.
29. Lee, E. and Nass, C., Experimental Tests of Normative Group Influence and Representation Effects in Computer-Mediated Communication: When Interacting Via Computers Differs from Interacting with Computers. *Human Computer Research*, 2002. 28(3): p. 349-381.
30. Marti, S. and Schmandt, C. Physical Embodiments for Mobile Communication Agents. in *UIST*. 2005. Seattle, Washington.
31. McBreen, H. and Jack, M., Evaluating humanoid synthetic agents in E-retail applications. *IEEE SMC Transactions*, 2001. 31: p. 394-405.
32. McBreen, H. and Jack, M., Evaluating Humanoid Synthetic Agents in E-Retail Applications. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Stystems and Humans*, 2001. 31(5): p. 394-405.
33. Moreno, R., Mayer, R., Spires, H., and Lester, J., The case for social agency in computer-based teaching: Do students learn more deeply when they interact with

animated pedagogical agents?. *Cognition and Instruction*, 2001. 19: p. 177-213.
34. Moreno, R., Mayer, R., Spires, H., and Lester, J., The Case for Social Agency in Computer-Based Teaching: Do Students Learn More Deeply When They Interact With Animated Pedogogical Agents? *Cognition and Instruction*, 2001. 19(2): p. 177-213.
35. Mori, M., The Uncanny Valley. *Energy*, 1970. 7: p. 33-35.
36. Mosteller, F. and Bush, R., Selected quantitative techniques, in *Handbook of social psychology: Vol. 1. Theory and method*, G. Lindzey, Editor. 1954, Addison-Wesley: Cambridge, MA. p. 289-334.
37. Moundridou, M. and Virvou, M., Evaluating the Persona Effect of an Interface Agent in an Intelligent Tutoring System. *Journal of Computer Assisted Learning*, 2002. 18(3): p. 253-261.
38. Murano, P. Effectiveness of Mapping Human-Oriented Information to Feedback From a Software Interface. in *24th International Conference Information Technology Interfaces ITI* 2002. Cavtat, Croatia.
39. Nass, C., Moon, Y., and Carney, P., Are respondents polite to computers? Social desirability and direct responses to computers. *Journal of Applied Social Psychology*, 1999. 29: p. 1093-1110.
40. Nowak, K. and Rauh, C., The influence of the avatar on online perceptions of anthropomorphism, androgyny, credibility, homophily, and attraction. *Journal of Computer-Mediated Communication*, 2005. 11.
41. Okonkwo, C. and Vassileva, J. Affective Pedagogical Agents and User Presuasion. in *Universal Access in Human-Computer Interaction*. 2001. New Orleans, USA.
42. Oulasvirta, A. and Salovaara, A., A cognitive meta-analysis of design approaches to interruptions in intelligent environments. *Proceedings of CHI 2004*, 2004: p. 1155-1158.
43. Prendinger, H., Ma, C., and Yingzi, J. Understanding the Effect of Life-Like Interface Agents Through Users' Eye Movements. in *International Conference on Multimodal Interfaces*. 2005. New York: ACM Press.
44. Prendinger, H., Mori, J., and Ishizuka, M., Using Human Physiology to Evaluate Subtle Expressivity of a Virtual Quizmaster in a Mathematical Game. *International Journal of Human-Computer Studies*, 2005. 62: p. 231-245.
45. Qvarfordt, P., Jonsson, A., and Dahlback, N. The Role of Spoken Feedback in Experiencing Multimodal Interfaces as Human-like. in *ICMI '03*. 2003. Vancouver, British Columbia, Canada.
46. Rickenberg, R. and Reeves, B. The Effects of Animated Characters on Anxiety, Task Performance, and Evaluations of User Interfaces. in *CHI*. 2000.
47. Rosenthal, R., Meta-analytic procedures for social research. 1984, Beverly Hills, CA: Sage Publications.
48. Rosenthal, R. and DiMatteo, M., Meta-Analysis: Recent Developments in Quantitative Methods in Literature Reviews. *Annual Review of Psychology*, 2001. 52: p. 59-82.
49. Schaumburg, H., Computers as Tools or as Social Actors? - The Users' Perspective on Anthropomorphic Agents. *International Journal of Cooperative Information Systems*, 2001. 10: p. 217-234.
50. Slater, M., How colorful was your day? Why questionnaires cannot assess presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 2004. 13: p. 484-493.
51. Sproull, L., Subramani, R., Kiesler, S., Walker, J., and Waters, K., When the Interface Is a Face. *Human-Computer Interaction*, 1996. 11: p. 97-124.
52. Swartout, W., Gratch, J., Hill, R., Hovy, E., Marsella, S., Rickel, J., and Traum, D., *Toward Virtual Humans*, in *AI*. 2006.
53. Takeuchi, A. and Naito, T., Situated Facial Displays: Towards Social Interaction. *Proceedings of CHI 1995*, 1996: p. 450-455.
54. Tomlinson, B., Yau, M., and Baumer, E. Embodied Mobile Agents. in *Fifth International Joint Conference on Autonomous Agents and Multi Agent Systems*. 2006. Hakodate, Japan.
55. Van Mulken, S., Andre, E., and Muller, J. The Persona Effect: How Substantial is It? . in *Human-Computer Interaction*. 1998. Berlin, Germany.
56. Van Mulken, S., Andre, E., and Muller, J. An Empiricial Study on the Trustworthiness of Life-Like Interface Agents. in *Human-Computer Interaction*. 1999. Mahway, New Jersey: Lawrence Erlbaum Associates.
57. van Vugt, H., Konijn, E., Hoorn, J., Keur, I., and Eliëns, A., Realism is not all! User engagement with task-related interface characters. *Interacting with Computers*, in press.
58. Vertegaal, R. and Ding, Y. Explaining Effects of Eye Gaze on Mediated Group Conversations: Amount or Synchronization. in *Conference on Computer Supported Cooperative Work*. 2002. New Orleans: ACM Press.
59. Walker, J., Sproull, L., and Subramani, R., Using a Human Face in an Interface. *Proceedings of CHI 1994*, 1994: p. 85-91.
60. Walker, J., Sproull, L., and Subramani, R. Using a Human Face in an Interface. in *CHI '94 "Celebrating Interdependence"*. 1994. Boston, Massachusetts USA.
61. Walther, J., Anderson, J., and Park, D., Inpersonal effects in computer-mediated interactions: A meta-analysis of social and antisocial communication. *Communication Research*, 1994. 21: p. 460-487.
62. Walther, J., Slovacek, C., and Tidwell, L., Is a Picture Worth a Thousand Words? Photographic Images in Long-Term and Short-Term Computer-Mediated Communication. *Communication Research*, 2001. 28: p. 105-134.

63. Weisband, S. and Kiesler, S., Self-disclosure on computer forms: Meta-analysis and implications. . *Proceedings of CHI96*, 1986.
64. Wexelblat, A., Don't Make that Face: A Report on Anthropomorphizing an Interface. *Intelligent Environments*, 1998.
65. Xiao, J., Stasko, J., and Catrambone, R. Embodied Conversational Agents as UI Paradigm: A Framework for Evaluation. in *AAMAS02 Workshop on 'Embodied Conversational Agents - Let's Specifiy and Evaluate Them!* 2002. Bologna, Italy.
66. Zanbaka, C., Goolkasian, P., and Hodges, L., Can a virtual cat persuade you? The role of gender and realism in speaker persuasiveness. *Proceedings of CHI 2006*, 2006: p. 1153-1162.

## APPENDIX A – EFFECT SIZES AND SIGNIFICANCE VALUES OF STUDIES INCLUDED

| | **Performance** | | **Subjective** | | |
|---|---|---|---|---|---|
| | Face vs. No Face | High vs. Low Realism | Face vs. No Face | High vs. Low Realism | **N** |
| Okonkwo & Vassileva, 2001 [41] | | r = 0, z = 0.24 | | r = 0.03, z = 0.84 | 12 |
| Moundridou, Virvou 2002 [37] | r = 0.1, z = 0.39 | | r = 0.48, z = 4 | | 48 |
| Hongpaisanwiwat & Lewis, 2003 [23] | r = 0, z = -0.02 | r = 0.07, z = 0.45 | | | 50 |
| Burgoon, Bengtsson, Bonito, Ramirez, & Dunbar, 1999 [11] | r = 0.03, z = 0.2 | r = -0.03, z = -0.17 | r = 0, z = -0.04 | r = 0.12, z = 0.8 | 50 |
| Bailenson, Beall, & Blasovich, 2002 [2] | | | r = 0.51, z = 1.92 | r = 0.16, z = 0.46 | 30 |
| Burgoon, Bonito, Bengtsson, Cederberg, Lundeberg, Lundeberg, & Allspach, 2000 [12] | r = 0.04, z = 0.19 | r = -0.04, z = -0.2 | r = 0.06, z = 0.33 | r = 0.14, z = 0.57 | 50 |
| Bailenson, Blasovich, Beall, & Loomis, 2001 [1] | | r = 0.13, z = 1.32 | | r = 0.2, z = 1.97 | 50 |
| Vertegaal & Ding, 2002 [58] | | r = -0.08, z = -0.78 | | | 34 |
| Bailenson, Blasovich, Beall, & Loomis, 2003 [3] | | r = -0.11, z = -0.67 | | r = 0.42, z = 3.87 | 70 |
| Bente, Kramer, Petersen, & Ruiter, 2001 [7] | | r = -0.15, z = -1.33 | | | 100 |
| Bonito, Burgoon, & Bengtsson, 1999 [10] | | | r = -0.09, z = -0.39 | r = -0.07, z = -0.32 | 30 |
| Prendinger, Ma, & Tingzi, 2005 [43] | r = 0.39, z = 1.88 | | | | 20 |
| Gerhard, Moore, & Hobbs, 2005 [19] | | | r = 0.62, z = 2.22 | | 20 |
| Hook, Persson, & Sjolinder, 2000 [24] | | | r = 0.12, z = 0.71 | | 38 |
| Moreno, Mayer, Spires, & Lester, 2001 (a) [34] | r = 0.26, z = 1.85 | | r = 0.06, z = 0.39 | | 44 |
| Moreno, Mayer, Spires, & Lester, 2001 (b) | r = 0.22, z = 1.55 | | r = 0.15, z = 0.63 | | 48 |
| Moreno, Mayer, Spires, & Lester, 2001 (c) | r = 0.18, z = 1.27 | | r = 0.05, z = 0.37 | | 38 |
| Moreno, Mayer, Spires, & Lester, 2001 (d) | r = 0.17, z = 1.3 | r = -0.18, z = -1.36 | r = 0.18, z = 1.41 | r = 0.16, z = 1.29 | 64 |
| Moreno, Mayer, Spires, & Lester, 2001 (e) | r = 0.38, z = 2.16 | r = 0.12, z = 0.61 | r = 0.15, z = 0.85 | r = -0.09, z = -0.63 | 79 |
| Schaumburg, 2001 [49] | | | r = 0.29, z = 3.15 | | 105 |
| Walther, Slovacek, & Tidwell, 2001 (a) [62] | | | r = -0.2, z = -0.82 | | 14 |
| Walther, Slovacek, & Tidwell, 2001 (b) | | | r = 0.15, z = 0.6 | | 14 |
| Chan & Khalid, 2003 [15] | r = 0.24, z = 1.71 | | | | 48 |
| Hess, Fuller, & Matthew, 2006 [22] | r = 0.03, z = -0.25 | r = -0.06, z = -0.86 | r = 0.05, z = -0.65 | r = 0.004, z = -0.5 | 180 |
| Kiesler, Waters, & Sproull, 1996 [25] | | | r = 0.23, z = 1.61 | r = 0.36, z = 2.18 | 18 |
| Cowell, 2001 [16] | | | r = 0.24, z = 1.56 | | 36 |
| Lee & Nass, 2002 (a) [29] | | | r = 0.15, z = 1.04 | r = 0.18, z = 1.28 | 48 |
| Lee & Nass, 2002 (b) | | | r = -0.11, z = -0.56 | r = 0.16, z = 1.16 | 48 |
| Bailenson, Swinth, Hoyt, Persky, Dimov, & Blascovich (2005) [4] | | | r = 0.16, z = 2.38 | | 210 |
| Guadagno, Blascovich, Bailenson, McCall (in press) [20] | | | | r = 0.25, z = 2.58 | 100 |
| Van Vugt, Konijn, Hoorn, Keur, Eliëns (in press) [57] | | | r = .03, z = .24 | r = -.01, z = -.13 | 140 |
| **N** | 12 | 11 | 22 | 15 | |
| **Overall** | r = .14<br>$r^2$ = .02<br>z = 2.74<br>p = .006 | r = -.04<br>$r^2$ = .002<br>z = -1.46<br>p = .14 | r = 0.13<br>$r^2$ = .02<br>z = 4.83<br>p < .001 | r = 0.11<br>$r^2$ = .01<br>z = 2.79<br>p = .002 | |

**Note**: Details of the meta-analysis, including underlying aggregated means and standard deviations and how different conditions were combined in specific studies, are available at http://vhil.stanford.edu.

## APPENDIX B – RESULTS OF INFORMAL DESCRIPTIVE ANALYSIS

| | Subjective | | Behavioral | |
|---|---|---|---|---|
| | Face vs. No Face | High vs. Low Realism | Face vs. No Face | High vs. Low Realism |
| Okonkwo & Vassileva, 2001 [41] | | High | | High |
| Koda & Maes, 1996 [26] | Yes | High | | |
| Van Mulken, Andre, & Muller, 1999 [56] | No | High | Yes | High |
| Xiao, Stasko, & Castrambone, 2002 [65] | Yes | High | | |
| Bartneck, 2001 [6] | | High | | |
| Moundridou, Virvou 2002 [37] | Yes | | Yes | |
| Hongpaisanwiwat & Lewis, 2003 [23] | | | Yes | High |
| Burgoon, Bengtsson, Bonito, Ramirez, & Dunbar, 1999 [11] | Yes | High | Yes | Low |
| Bailenson, Beall, & Blasovich, 2002 [2] | Yes | High | | |
| Walker, Sproull, & Subramani, 1994 [60] | No | | Yes | |
| Sproull, Subramani, Kiesler, Walker, & Waters, 1996 [51] | No | High | | |
| Burgoon et al. 2000 [12] | Yes | High | - | Low |
| Wexelblat, 1998 [64] | No | | | |
| Murano, 2002 [38] | | High | | |
| Bailenson, Blasovich, Beall, & Loomis, 2001 [1] | | High | Yes | High |
| Vertegaal & Ding, 2002 [58] | | | | High |
| Fabri, Moore, & Hobbs, 2002 [18] | | | No | |
| Bailenson, Blasovich, Beall, & Loomis, 2003 [3] | | High | | - |
| Bente, Kramer, Petersen, & Ruiter, 2001 [7] | | | | Low |
| Van Mulken, Andre, & Muller, 1998 [55] | Yes | | - | |
| Bonito, Burgoon, & Bengtsson, 1999 [10] | No | - | | |
| Qvarfordt, Jonsson, & Dahlback, 2003 [45] | Yes | High | | |
| Rickenberg & Reeves, 2000 [46] | Yes | Low | Yes | Low |
| Koda, 2003 [27] | Yes | High | | |
| Marti & Schmandt, 2005 [30] | | High | | |
| Prendinger, Ma, & Tingzi, 2005. [43] | | | Yes | |
| Prendinger, Mori, & Ishizuka, 2005 [44] | Yes | | | |
| Gerhard, Moore, & Hobbs, 2005 [19] | Yes | | | |
| Gulz, 2005 [21] | Yes | | | |
| Hook, Persson, & Sjolinder, 2000 [24] | Yes | | | |
| Kramer, 2005 [28] | | | Yes | High |
| McBreen & Jack, 2001 [32] | No | High | | |
| Moreno, Mayer, Spires, & Lester, 2001 [34] | Yes | | Yes | |
| Shaumburg, 2001 [49] | Yes | | | |
| Walther, Slovacek, & Tidwell, 2001 [62] | - | | | |
| Chan & Khalid, 2003 [15] | Yes | | | |
| Hess, Fuller, & Matthew, 2006 [22] | Yes | Low | Yes | High |
| Kiesler, Waters, & Sproull, 1996 [25] | - | High | | |
| Cowell, 2001 [16] | | | Yes | |
| Lee & Nass, 2002 [29] | Yes | High | | |
| Zanbaka, Goolkasian, Hodges (2006) [66] | | High | | |
| Beun, Vos, Witteman (2003) [8] | | | Yes | High |
| Bailenson, Swinth, Hoyt, Persky, Dimov, & Blascovich (2005) [4] | Yes | | | |
| Guadagno, Blascovich, Bailenson, McCall (in press) [20] | | | | High |
| Van Vugt, Konijn, Hoorn, Keur, Eliëns (in press) [57] | | | - | High |
| **N** | 28 | 22 | 17 | 15 |
| **Overall** | Yes = 20<br>No = 6 | High = 19<br>Low = 2 | Yes = 13<br>No = 1 | High = 10<br>Low = 4 |
| **%** | 71% | 86% | 76% | 66% |

**Note**: The conditions (i.e., "Yes", "No", "High", "Low") in the cells refer to the condition that resulted in more positive effects for that study. Thus, a "High" means that the high-realism condition out-performed the low-realism condition.