Effects of Femininity and Masculinity in Voice and Appearance on Stereotype Judgement

Christoffer Andersson
fpr10can@student.lu.se

Earlier studies have shown that both gender and masculine or feminine traits affects judgements about a person's qualities. This study builds upon earlier research and investigates the relation between voice and appearance with respect to gender based stereotype judgements, by letting participants listen to and then rate two of totally eight different digital characters, four female and four male. The manipulated variables were the femininity or masculinity of the characters' voices and the femininity or masculinity of their appearance. The ratings regarded the characters perceived competence, intelligence, and empathic abilities. The hypothesis that masculinity would increase ratings for competence and intelligence was unsupported for both male and female characters. Instead the results show that when female characters had matching voice and appearance (both feminine or both masculine), the characters was perceived as both more intelligent and competent.

Furthermore, the second hypothesis stated in the study, namely that feminine traits would increase perceived emphatic abilities, found support in the data, however only when comparing female characters.

The conclusions drawn from these findings are that it is plausible that common stereotypes might have lost some of its strengths, at least for the population studied and in the domain of medical doctors. Furthermore, the author suspects a matching effect. However, a more extensive study, incorporating a more complex and larger set of stimuli, is required in order to draw more generalised conclusions.

Keywords: gender, stereotype, matching effect, voice, appearance, intelligence, competence, empathy.

1 Introduction

The purpose of this paper is to compare the effect of visual and auditory traits with respect to gender based stereotype judgements. Earlier research has investigated how gender, as well as femininity and masculinity, affects judgements in a way that conforms to common gender based stereotypes, e.g. judging a character's intelligence or empathic competence, where males are often judged to be more competent and intelligent whilst women are judged to be nicer and more empathic (see e.g. Reeves & Nass, 1996; Nass & Brave, 2005; Gulz, Ahlner & Haake, 2007). This has previously been studied by altering either visual (Gulz, Ahlner& Haake, 2007) or auditory (Reeves & Nass, 1996) traits, but there seems not to be any trace in the literature of an experimental setup which investigates both. This leaves a gap of knowledge about which (if any) of the cues affect gender based stereotypes the most, or how they may interact. This study will therefore investigate the interaction of visual and auditory stimuli.

To do this four virtual characters were used, two female and two male (digital characters borrowed from Gulz, Ahlner & Haake, 2007, see Fig. 1) that were designed to look either masculine or feminine by altering shoulder width and head characteristics, such as protrution of jaw and nose, and overall colour scheme. The visual stimuli in this study were also given different voices, two male and two female versions, digitally altered from one male and one female voice to sound either masculine or feminine by changing the voices' fundamental frequency. Participants listened to an informational speech from two of the virtual character, and was after each one of these speeches asked to answer questions regarding the character's competence, intelligence, and empathy on an eight point Likert scale.

In other words, this study investigates how stereotype judgements is affected by masculinity and femininity manipulated via both auditory and visual traits.

What is a Stereotype?

Throughout the literature there has been no consensus on how to define the stereotype concept. On the one hand there are attempts to define the stereotype concept so that it includes the negative connotations usually attributed to it. For example, Haake and Gulz (2008) proposes that a stereotype is to be defined in contrast to prototypes. Prototypes are a neutral representation of and, as they write, “a 'typical exemplar' of a concept” (ibid., p. 3), whereas a stereotype brings with it negative connotations. These negative connotations are closely related to the notion that a stereotype is an inaccurate way of inferring attributes to single members of a certain group, or from a certain person to a larger group. Furthermore, it should be pointed out that according to this viewpoint the negative connotations here are attributed to stereotyping in and of itself and not contingent on whether the stereotype infers negative traits. For example, a neutral prototype of a serial killer can still bring about negative connotations, and a racist stereotype of Asians being inherently good at mathematics is still a stereotype that infers a positive trait to a certain group.

On the other hand there are attempts to avoid defining stereotypes as something negative in and of itself. For example, Hilton and von Hippel (1996) writes: “[W]e adopt the standard viewpoint that stereotypes are beliefs about the characteristics, attributes, and behaviors of members of certain groups. More than just beliefs about groups, they are also theories about how and why certain attributes go together” (ibid., p. 240). They go on by pointing out that “stereotypes are sometimes accurate representations of reality, […] or at least of the local reality to which the perceiver is exposed” (ibid.). This is a way of defining stereotypes that instead of contrasting it with, makes them a subset of prototypes.

Another argument by Hilton and von Hippel (1996) for defining the stereotype concept as a subset of prototypes that may very well be neutral and accurate is that we already have a word for unreliable and inaccurate inferences, namely “pre-

judices". However, in the colloquial usage of "stereotype" and "prejudice" there is a difference in meaning that would render both words useful. Prejudices tends to refer to personal and individually held beliefs about certain groups, whilst stereotypes refers to beliefs that are widespread and socially nurtured. The definition then, that keeps the negative connotations that the stereotypes brings with it, seems to the author of this study at least, to be the most attractive one.

Stereotype Research

The research concerning gender based stereotypes has shown how widespread and seemingly fundamental this phenomena is. One study, conducted by Nass and Brave (2005), has even shown that people exhibit stereotypical judgements when listening to texts from eBay auctions, knowing that they were read by actors unaffiliated with the authors. When texts from auctions that were verified to be in male domains (e. g., books about guns) were read by a man, the participants judged the description to be more credible and the author more competent than when the same text was read by a woman. However, the results also showed that in tests regarding a typical feminine domain, such as sewing, the female voice was regarded more competent.

In another study, by Nass & Brave (2005), which was primarily conducted to investigate *similarity attraction* (that persons are more prone to liking other persons of the same gender as themself), participants read about different dilemmas on a computer screen. The participants then clicked a button to get advice about the dilemma from a synthetic voice, which would be either male or female, that argued for one of two solutions to the dilemma. Participants could then, on an eight point Likert scale, state which way they were leaning regarding the two solutions and how much they liked the advice given. The "similarity attraction"-hypothesis was confirmed. Men liked the male voice more than the female voice and women liked the female voice over the male version. Interestingly they also found support for a post hoc hypothesis. The preferred solutions to the dilemma conformed with the advice given by the male voice more than what they did with the female counterpart. That is, the male advice had a greater effect on both male and female participants' preferred solutions. These results are quite surprising when taking into account that the participants in both experiments knew that the voice reading the description had no connection with text production and they could also hear that the voice was synthetic. This indicates that gender plays a fundamental role even when we know it should not. Furthermore, stereotyping seems to be a non-reflective, subconscious trait rather than a product of deliberate judgements. Moreover, results like these suggest that stereotypes are well rooted in our everyday decision making.

These widespread stereotypical judgements do not only affect judgements about others, but seem to also affect the the thoughts of oneself. In another study, by Cadinu et al. (2005), female participants were asked to perform a mathematical test. Before the test the control group was told that there was no difference between gender in regards to mathematical ability, whilst the test group was told that women had been shown to perform poorer than males. The test group that had been exposed to what Cadinu et al. (2005) called the "stereotypical threat" performed poorer than the control group. These findings seem to show the possible negative effects of stereotypes in the element of social feedback when evaluating oneself. To assign oneself to a stereotypical category can thus seemingly override even self knowledge.

Moreover, gender based stereotypes can also apply to femininity and masculinity as well, which seems to be a fur-

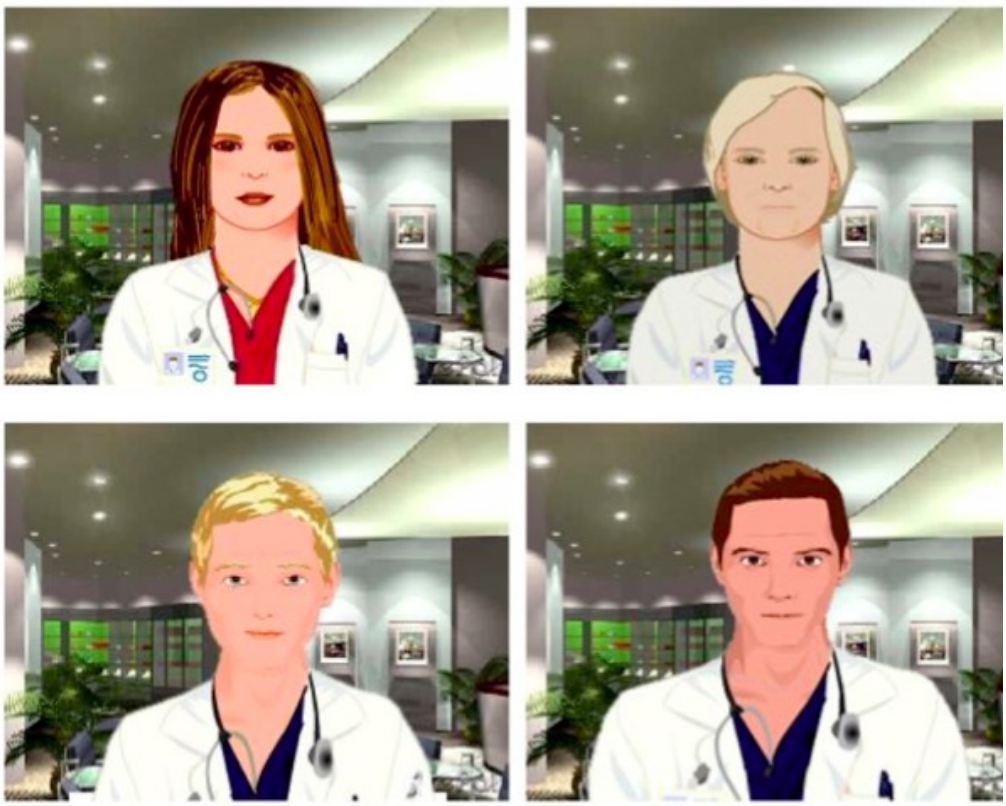

Fig. 1. *Stimuli used in Gulz, Ahlner, & Haake's (2007) study (used with permission). Upper left shows the female feminine character, the upper right shows the masculine female character, the lower left shows the feminine male character, and the lower right shows the masculine male.*

ther distinction that also plays a role in how we subconsciously assess a person's properties, e.g. competence. Research that focuses on masculinity and femininity shows that the subjects not only assess men to be more intelligent, but furthermore that the more masculine that person's traits are, the more intelligent the person is thought to be (see e.g. Reeves & Nass, 1996). In another study concerning visual cues of masculinity and femininity, Gulz, Ahlner and Haake (2007) let participants listen to virtual doctors explaining the pros and cons of working night-shifts, and then answer a multitude of questions regarding the virtual doctor's personality traits and dispositions. The virtual doctors were male or female, with either feminine or masculine visual traits (see Fig.1). The female voice was the same for both female characters and the male voice was the same for both male characters. However, a full evaluation of result was hindered due to an unexpected variable in that the male and female voice differed in dialects, which Gulz, Ahlner and Haake (2007) discussed might have conflicted with the gender's effect on participants' judgements. However, when comparing the results of the two female characters the results seemed to conform with previously known stereotypes, that is association between masculinity, intelligence, and competence on the one hand, and between femininity and empathy on the other.

In a similar study by Reeves and Nass (1996), female voices were altered into sounding either masculine or feminine, and were accompanied by pictures of six different virtual female characters. Participants were exposed to six different settings, three with feminine and three with masculine female voices. The results were in line with other findings in similar studies, and the more masculine female voices were assessed as both more intelligent and persuasive.

Finally, it certainly seems like gender stereotypes are a widespread, robust, and sub-conscious phenomena, that are hard to overcome even when using reflective and explicit reasoning about genders role in assessing other individuals properties. In the next section I will try to link the discussed findings, and give a theoretical account of the *stereotype* concept in another field of research, namely decision making theory.

Possible Mechanics for Stereotype Judgements

Stereotypes may be seen in the light of theories in decision making research. Usually stereotypical judgements can be described as wrongfully inferring that one personal quality (e.g. gender, race, or weight) entails other qualities of that person (e.g. intelligence). However, stereotypes are often not explicit, and persons that say they neither like or condone common stereotypes may unknowingly exhibit them, as has been shown by implicit association tests (see Latu et al., 2011; Steffens & Jelenec, 2011; Sabin, Marini & Nosek, 2012; Blair et al., 2013). This suggests that stereotypes need not to be the product of a deficiency in moral competence or explicit reasoning skills, but rather something implicit and automatic, which may be explained by a certain field of decision making theory, namely heuristics research.

Decision making research as a whole can be divided into two main fields, where one is mostly concerned with how decision making should be made by ideal decision makers, whilst the other is concerned with describing peoples actual decision making strategies, and take the limits of human computational and epistemological capabilities into consideration when formulating ideas of the rationality of decision strategies. The latter field of research has been trying to both explore and vindicate the effectiveness (in regards to precision, cognitive load, as well as decision speed) of simple heuristics, that is, simple decision rules. Investigation in the area of heuristics is conducted mostly by modelling possible decision rules and running them through computer simulations, and then evaluating their effectiveness, often compared with more advanced algorithms from the field of ideal decision making. This approach does, of course, not answer whether the particular decision rule is common, or even used by real people. However, other studies indicate that most of our everyday decision making is made in a fast, frugal, albeit sub-conscious manner, as heuristic research predicts (Gigerenzer, Todd, & the ABC Research Group, 1999; Kahneman, 2011).

In one famous study, described in detail by Goldstein and Gigerenzer (2002), American students had to choose which of two cities had the larger population. The pair of cities where either American or German, which would make one think that the American students made better judgements about American cities than of foreign, not so well known, German ones. However, the results showed that the American students made equally, or even slightly better, when choosing between German cities. This was interpreted as a result of using a simple decision rule called the *Recognition Heuristic*, where lack of knowledge actually can benefit the decision maker. When recognising just one of the two presented German cities, the students would choose the known city, which more often than not was the larger city. Cities that have a larger population tend to be presented in media and everyday conversations more frequent than smaller cities, which makes recognition a somewhat precise indicator of population size, and since the students probably recognised almost all of the American cities, the *Recognition Heuristic* would be useless. More information would have to be computed, as would knowledge about what information had a high predictive value of population size.

This process, when the decision maker is going to decide between two known stimuli, has been explored by heuristics research. One of the most well described and developed heuristics is the *Take the Best* algorithm (see Fig. 2), which is meant to handle situations where the *Recognition Heuristic* is

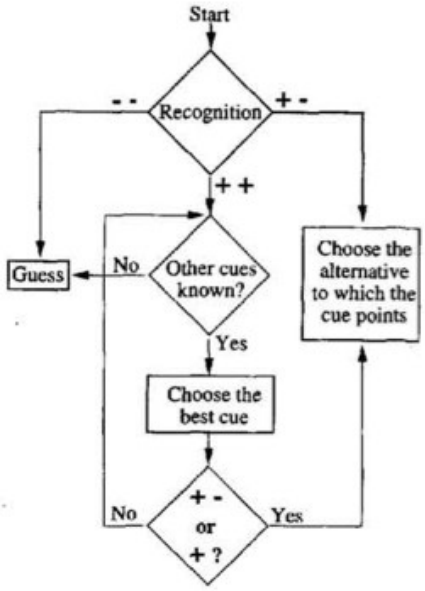

Fig. 2. *The Take the Best algorithm from Gigerenzer & Goldstein (1996) (used with permission).*

bypassed by knowing more than one of the stimuli. The *Take the Best* (TTB) algorithm is operated by going trough the knowledge base about the known stimuli and then choosing the discriminating cue that is thought to best predict a correct choice. For example, when choosing between two known cities, you might think that one of them probably is larger since it has a famous sports team whilst the other city does not.

To test the effectiveness of the TTB algorithm, Gigerenzer and Goldstein (1996), simulated a decision making environment and let a multitude of algorithms, most of them constructed by the normative camp, run through the simulation. The algorithm's task was the same as the former discussed American students, to decide which of two German cities had the largest population. The simulation was run through many settings varying both in degree of recognised cities (based on American surveys on the most commonly recognised German cities) and knowledge of the recognised cities' cues. When evaluating the results it seemed that the TTB heuristic did as good as, or better than, all other algorithms when it came to precision. Since the algorithms from the normative camp always used all available information and also weighted all cue values into one final value for the the two options, they would be much slower if used by a real decision maker with limited computational capacity. The conclusions seem then be that the TTB is both fast and frugal without having to compromise with precision (Gigerenzer & Goldstein, 1996).

However, during the simulation the TTB was as unbounded as the algorithms from the normative camp regarding one aspect, which may account for some of the TTB's success. When both cities were recognised, the TTB searched the cities' cues to choose the one with the highest ecological value, that is the highest predictive power. If the cue discriminated the options it was chosen, and if it did not discriminate, next cue was searched. This means that the ecological value was programmed as a part of the environment, with no subjective barriers to simulate the limitations of real decision making agents. Surely, a real decision making agent has no objective and numerically precise list of cues to be used to predict the precision of a choice. Instead another option is available to understand the heuristic's cues.

Contrary to the idea of ecological cues and values (used by Gigerenzer & Goldstein, 1996), which assumes that the decision maker have a near perfect understanding of frequencies and the precision of inferences, a more subjective approach can be found in Tversky & Kahneman's (1973) research. Here, inferences are made by the availability, that is the ease of retrieval, of cues. For example, one of their studies showed that people often misjudge the frequency of words that begin with a certain consonant (e.g. *k*) compared to when the consonant is the third letter of a word. Words that starts with a *k* was assessed as being more frequent, even though they are not. Tversky and Kahneman (1973) suggest that this is because the availability is greater for words that start with a *k* compared to the availability of words where *k* is the third letter. Other findings, such as exaggeration of the frequencies of e.g. tornadoes occurring (Gigerenzer et al., 1999, p. 214) seem to also point to the conclusion that the assumption of neat ecological values in cue hierarchies are misguided.

Gender based stereotypes, then, may be seen in the light of this part of decision making research. When inferring a person's qualities, such as intelligence, one might have greater availability of men working in high status, intelligence requiring work, which in turn can be seen as a consequence of gender based stereotyping. For example, when hiring a new doctor, the employer might subconsciously judge men differently than women, since the employer might have more frequent experiences with male doctors. This might cause the gender cue to rise in the inner hierarchy of important cues that will be selected for inferring a person's competence as a doctor. Then, when making a decision, not only is the gender cue too high (in comparison with its actual, objective, predictive value) in the hierarchy, but also will the availability of competent male doctors be greater than that of competent female doctors.

This is then a way in which stereotypical judgements can be understood. It could be argued that, when the predictive power of a person's quality *x* is overvalued (e.g. due to inaccurate social information) in a subject's hierarchy of cues, and thus is used to infer the person's other qualities, a stereotypical judgement has occurred. Thus gender based stereotypes entails that with a person's gender, other qualities follows.

However, from 2005 to 2012 the proportion of female medical doctors in Sweden has risen from 42.3% to 47.7% (SCB, 2013), which is concern of this study since the characters used in this study as stimuli are all portrayed as medical doctors. If the equalising distribution of male and female doctors have been perceived by the participants, then this might interfere with otherwise common gender based stereotypes.

Main focuses

Earlier research has until now shown that both visual cues in a person's appearance, and certain properties of a person's voice may elicit assessments that conform with gender based stereotypes (Gulz, Ahlner & Haake, 2007; Reeves & Nass, 1996). Three qualities seen to vary when manipulating gender and masculinity/femininity were intelligence, competence and empathy. The latter was rated higher for women, and was also enhanced with feminine traits, while the two former were assessed as higher in male and was similarly enhanced by masculine traits in persons with any gender. Thus, two clear hypotheses may be formulated:

Hypothesis 1: Stimuli with masculine voice and appearance will be assessed as more intelligent and competent than stimuli with feminine voice and appearance.

Hypothesis 2: Stimuli with feminine voice and appearance will be assessed as more empathic than stimuli with masculine voice and appearance.

Furthermore, a possible comparison of the visual and auditory traits opens the possibility for questions that have not yet been tried or answered. Former research has addressed how singular masculine and feminine traits in males and females affects stereotype assessment, but not what happens when both voice and appearance are altered. This is thus unknown. For example, which of a female with masculine voice and feminine appearance and a female with feminine voice and masculine appearance will elicit higher degree of stereotype assessment?

Another interesting question that arises from this possibility to investigate if any of the traits are dominant, is to see if this dominance is the same for both genders. While one of the traits, e.g. the voice, dominates stereotype judgements for the female stimuli, the appearance might dominate stereotype judgements for the male stimuli. Since these questions have

been, to the author's knowledge, unexamined, no predictions will be made. Instead they are formulated as open and exploratory questions.

Question 1: Is any of the two traits (appearance and voice) more dominant than the other?

Question 2: If dominating traits are found, are they the same for both genders?

2 Methods

Visual stimuli

The visual stimuli used for this study was used with permission from the Gulz, Ahlner and Haake's (2007) study that investigated visual stereotypes both within gender and the femininity to masculinity variable. The variables altered in the forming of these stimuli was described in the following way:

"- Feminine character (F+): Manipulated with feminine attributes such as: the *baby- face* scheme (rounded head shapes, bigger eyes, smaller nose, narrower shoulders); long (colored) hair and make up, that pronounces feminine attributes by enlarging the eyes, making them rounder and more distinct and making the lips fuller.

- *Weak feminine character (F-):* Manipulated with masculine attributes such as: broader head, a more angular and pronounced jaw, a high for-head; paler colors as to eyes, mouth and hair, which weakens the impact of these female attributes; overall paler color scheme reducing the number of distinct features and thus weakening any categorization of gender ! whether feminine or masculine.

- *Weak masculine character (M-):* Manipulated with feminine attributes such as: rounder and less pronounced shapes of head, jaw and nose; narrower shoulders; slightly red lips in combination with an overall paler color scheme, that weakens any distinct categorization of gender.

- *Masculine character (M+):* Manipulated with masculine attributes such as: broader, angular and more pronounced head shapes; broader shoulders, a distinct Adam's apple, pronounced, dark eye brows; neatly done hair; a more prominent colour scheme which produces distinct features and strengthens the categorization with respect to gender." (Gulz, Ahlner & Haake, 2007, pp. 658-659)

To control for any effect of the different colour schemes of the masculine version of the female characters (see Fig. 2), this version was given a slightly darker skin colour for this study. Previous validation of these stimuli, by Gulz, Ahlner and Haake (2007), will still be deemed as satisfactory.

However, the virtual characters are all dressed like doctors, which might have an effect of raising the overall competence and intelligence ratings. Such affect on ratings should however, a priori, be equal across the four virtual characters.

Auditory stimuli

The auditory stimuli made use of one male and one female recorded voice, which was altered on only one variable, the fundamental frequency. A typical male voice have a fundamental frequency of 128 Hz, but can range between 85-196 Hz, while the typical female voice is 225 Hz, but can range between 155-334 Hz (Williamson, 2006, p. 177).

The feminine female voice was altered to 240 Hz, that is 15 Hz from the female mean, while the masculine female voice was altered to 205 Hz, which is 20 Hz away from the mean value. The difference in manipulation compared to the female mean value was due to consideration of the perception of the different voices. The same kind of manipulation was done to the male voice, where the masculine version was lowered to 115 Hz, that is 13 Hz from the mean value of 128 Hz. The feminine version was altered to 142 Hz, that is 14 Hz above the mean value for men. This is supposed to be quite a subtle change, and is far from any extreme values, which makes comparison between appearance and voice easier to perform, since the appearance of the characters have been validated to be subtle in the Gulz, Ahlner and Haake (2007) study. The feminine and masculine versions of the male and female voices were validated using a convenience sample of 14 participants with a mean age of (approximately) 28 years, which were asked how masculine or feminine they thought the voices sounded on a seven point Likert scale. The results followed expectations and the mean values of the ratings is the following (the first letter, F or M, refers to gender, while the second letter, F or M, refers to feminine or masculine voice): FF = 1.93; FM = 4.71; MF = 2.81; MM = 6.50. The sample is however too small to make any further statistical analyses, but there descriptive results indicated a trend that conformed with expectations.

The text presented by the characters regards information and advice about working night shifts, and has been validated to be gender neutral both concerning the specific domain and word usage (Gulz, Ahlner & Haake, 2007).

This study makes use of the same voice recordings as did the Gulz, Ahlner and Haake (2007) study. Since they noticed a clear effect from the different dialects in the male and female voices, this study will not make any inter-gender comparisons. Instead results will only be analysed within the characters gender. However, it might at this point be important to point out that this study does not try to perform any conclusive comparison between the effect of voice and appearance, but aims instead to give a first look into any interaction or dominance between visual and auditory traits.

Procedure

This study utilised two ways of gathering data. One of these was through manually asking people around public gathering areas (libraries and study halls) to participate in the study. The participants were all told that the experiment took 5-8 minutes and was a part of the authors master's thesis in Cognitive Science. If accepting to participate, the subject was then given a laptop with the experiment presentation screen, and a pair of headphones to make the character presentation audible. They were then told to follow the instructions. After every session the experimenter asked if the participant had any questions surrounding the experiment or the master's thesis.

The second way of data gathering was through spreading a link to the experiment in social media forums. The experiment was identical to the one used in manual data collection. Since there was a concern that some participants might start doing the experiment in a noisy environment, or without sufficient audio systems to get a clear and audible sound from the characters' presentation, this was explicitly written on the presentation screen as a precondition for doing the test.

The Experiment

The experiment was programmed using web application built in PHP, and used a private web server to gather the data on to a simple spreadsheet. The experiment consisted of 7 pages, where the first was presented the experiment as a part of a

master's thesis in Cognitive Science. It also informed the participants that they needed to use earphones or sit in a quite environment to continue. Finally this text let the participants know that continuing from this page implied an agreement, and to allow the data gathered from the experiment to be used in the study.

The second page asked the participants provide age and gender information. The third page contained the first of two digital characters. The experiment was preprogrammed to make sure that the pairs of characters that each participant encountered did not differ in gender, and always differed in the two traits, so if the first character was a male with feminine voice and appearance, the second would be male with masculine voice and appearance. The presentation was manually started using a play button and could also be paused using a stop button. On the top of the page instructions were given to start the presentation and continuing when it was done. The fourth page was a nine item questionnaire, where three items each targeted the three categories discussed above (intelligence, competence, and empathy) on an eight point Likert scale. One of the three items in every targeted category was a negation, to make it possible to evaluate if the participants answered the questions thoughtfully. This item will therefore only be used as a control and not be a part of later analyses, since participants may vary on their sensitivity to framing effects (see e.g. Tversky & Kahneman, 1986). One of the three items in every category used the targeted word in the question, while others was formulated using synonyms. Page 5 and 6 presented the second character in the pair and was otherwise identical. The last page thanked the participant for taking part of the experiment

Every participant thus listened to and rated two characters. The characters was paired together by gender so that participants never rated one male and one female character, to decrease the risk of the participants to understand that the experiment investigated gender based stereotypes. The characters were also paired so that no traits were shared between them. This means that the pairs the participants could be presented with was either FMM & FFF, FMF & FFM, MMM & MFF, and MMF & MFM. The order of which of the two characters that was presented first did also alter.

Participants and data comparison

Data from 233 participants was collected. However, 25 of the 465 (or 5.4%) rows of data were deleted and not used in further analyses after checking for discrepancies in the participants' answers. If participant's answers had a standard deviation over 2,1 units in all domains (competence, intelligence, and empathic abilities) they were excluded from further analyses. This filtering of data was to account for, and detect linguistic misunderstandings and mistakes. For example, one participant from the web based version of the experiment had given the ratings 8, 8, and 1 for each of the three domains, which practically means that the participant had clicked the furthest to the right on each of the answers. Since one of the three questions of every domain was a negation of a synonym of the key word, and thus gave the opposite amount of points (8-1 instead of 1-8), these discrepancies could be found. Likewise, answers that generated high standard deviations shows that some kind of mistake or misunderstanding of how the words in the questionnaire were meant to be understood. If one thinks both that a person is very intelligent but not smart at all, then these words are used different by the author and the participant, and the answers to these questions measure something else than what was meant to be measured.

However, data from 221 participants remained (135 females and 86 males). Participants age spanned from 18-69 with a median age of 26 (*IQR* = 6).

3 Results

Hypothesis 1

This hypothesis stated that stimuli with masculine voice and appearance will be assessed as more intelligent and competent than stimuli with feminine voice and appearance. As seen on Fig. 3 and Fig. 4, no such difference was found. The female character with two feminine traits (FFF) had an average competence rating of 6.67 (*SD* = 1.55), and the female character with two masculine traits (FMM) averaged 6.69 (*SD* = 1.42). The female characters with mixed traits, FFM and FMF, averaged a competence rating of 6.27 (*SD* = 1.24) and 6.02 *(SD* = 1.68) respectively.

When comparing the female characters FMM differed significantly from FFM and FMF in competence ratings. FFF differed significantly from FMF, and showed a significant trend of difference from FFM. FFM did not differ significantly from FMF (see Table 1).

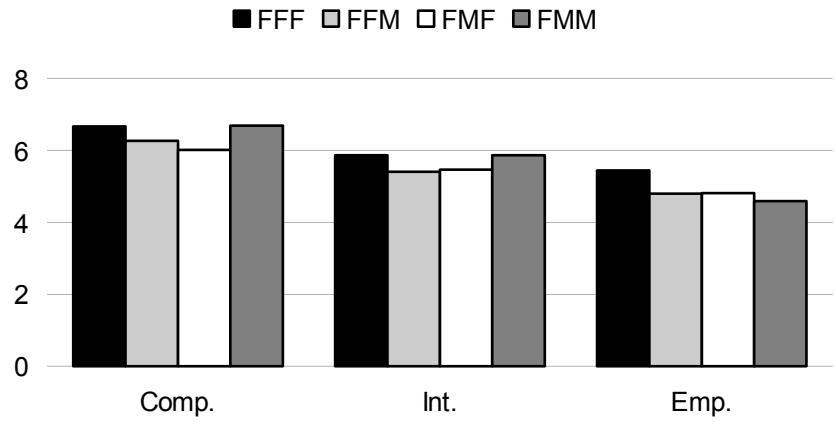

Fig. 3. *Rating averages for female characters. The first letter (F) refers to the characters' gender, the second (F or M) refers to appearance, and the third (F or M) refers to voice.*

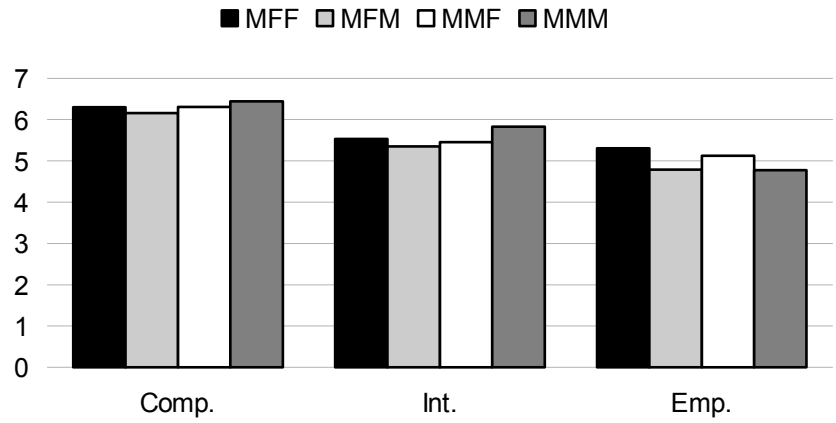

Fig. 4. *Rating averages for male characters. The first letter (M) refers to the characters' gender, the second (F or M) refers to appearance, and the third (F or M) refers to voice.*

For male characters the average competence ratings were the following: MFF = 6.30 (*SD* = 1.55), MFM = 6.16 (*SD* = 1.17), MMF = 6.31 (*SD* = 1.18), and MMM = 6.44 (*SD* = 1.32). Two-tailed T-tests comparing the ratings showed no significant differences in competence rating between any of the male characters (see Table 1).

The female characters averaged the following in intelligence ratings: FFF = 5.87 (*SD* = 1.53), FFM = 5.41 (*SD* = 1.35), FMF = 5.46 (*SD* = 1.38), and FMM = 5.87 (*SD* = 1.49). The same trend as in the competence ratings were found when comparing the ratings using two-tailed t-tests. FFF and FMM differed significantly from FFM and and showed a significant trend of difference to FMF, but not from each other. Neither did FFM and FMF differ significantly from each other (see Table 1).

The male characters' average ratings were: MFF = 5.53 (*SD* = 1.46), MFM = 5.35 (*SD* = 1.40), MMF = 5.48 (*SD* = 1.29), and MMM = 5.83 (*SD* = 1.50). The only pair of characters that differed significantly in their average ratings from each other when using two-tailed t-tests were MFM and MMM (see Table 1).

Hypothesis 2

The second hypothesis stated that stimuli with feminine

Table 1. *Shows the results of all two-tailed t-tests between the characters received ratings. The numbers written in underlined bold shows signifant results (p < .025), while numbers in bold without underlinings show significant trends (p < .05).*

Comp.	FFF	FFM	FMF	Comp.	MFF	MFM	MMF
FFM	**.035**	-	-	MFM	.436	-	-
FMF	**<u>.003</u>**	.214	-	MMF	.978	.339	-
FMM	.928	**<u>.021</u>**	**<u>.002</u>**	MMM	.475	.090	.418
Int.	FFF	FFM	FMF	Int.	MFF	MFM	MMF
FFM	**<u>.019</u>**	-	-	MFM	.357	-	-
FMF	**.042**	.765	-	MMF	.805	.461	-
FMM	1.000	**<u>.017</u>**	**.039**	MMM	.139	**<u>.015</u>**	.066
Emp.	FFF	FFM	FMF	Emp.	MFF	MFM	MMF
FFM	**.033**	-	-	MFM	.095	-	-
FMF	**.044**	.952	-	MMF	.635	.188	-
FMM	**<u>.007</u>**	.508	.482	MMM	.107	.958	.199

voice and appearance will be assessed as more empathic than stimuli with masculine voice and appearance.

The female characters' average ratings were: FFF = 5.45 (*SD* = 1.61), FFM = 4.80 (*SD* = 1.55), FMF = 4.81 (*SD* = 1.64), and FMM = 4.59 (*SD* = 1.71). Here, the all feminine character, FFF, differed significantly from FMM and showed a significant trend of difference to FFM and FMF (see Table 1).

The male characters received the following average ratings: MFF = 5.30 (*SD* = 1.69), MFM = 4.79 (*SD* = 1.50), MMF = 5.16 (*SD* = 1.47), and MMM = 4.77 (*SD* = 1.65). No significant differences could be found in any of the comparisons between average ratings (see Table 1).

Question 1 & 2

The first of the two exploratory questions in this study asked whether any of the two traits, voice and appearance, dominated the other in affecting the participants' ratings.

One way of figuring out whether one trait is more dominant than the other is to compare the average ratings for two characters within each gender that share one of the traits, but differ in the other (e.g. FFF and FFM), and then do the same for the other set of characters share characteristics of one trait but still differ in the other trait (in this case then, FMF and FMM). If voice has a main effect on ratings then both of the sets of characters compared should give the same increase-decrease in ratings when altering voice.

When comparing FFF with FFM, having a feminine voice seems to increase ratings (see Table 1.) with .3 points in competence ratings, .46 points in intelligence ratings, and .65 in empathy ratings. However when comparing FMF with FMM having a feminine voice decrease competence ratings with .67 and intelligence ratings with .41.

The same pattern emerge for appearance. When comparing FFF with FMF, having a feminine appearance significantly increased competence ratings with .65, intelligence ratings with .41, and empathy ratings with .64 points. When comparing FFM with FMM, having a feminine appearance significantly decreased competence and intelligence ratings with .42 and .46 points respectively.

When comparing male characters the only significant difference found was between the intelligence ratings of MMM and MFM. However, this result alone provides no answer to these questions.

4 Discussion

As stated in the introduction, earlier research suggest that masculinity would increase perceived competence and intelligence, while femininity would increase perceived empathic abilities. That is however not what this study found when investigating both appearance and voice. For the male characters in this study no stereotyping effect could be found, but for the female characters instead of ratings conforming with common gender based stereotypes about competence and intelligence, matching voice and appearance in femininity or masculinity seems to be what affect perception. However, female characters' average ratings indicate that matching feminine traits enhances perceived emphatic abilities, although having matching masculine trait did not affect ratings compared with the female characters of mixed traits.

What explanation can be found for these results? One hypothesis is that when traits are mismatched as in having different characteristics of voice and appearance, one being feminine and the other masculine, it is harder to categorise the person. This might lead to a prima facie judgement that is more sceptical in nature. However, this only explains why there were differences between mixed trait characters and characters where feminine and masculine traits were matched. It does not explain why no significant differences in competence and intelligence ratings were found between the female characters that had matching traits. Perhaps the gender based stereotypes suggested by earlier studies are not that common in the sample collected. How representative this sample is to a bigger population is hard to know without comparing these results to larger studies, with a broader and more geographically diverged sample.

Another way to explain these findings are the ways in which the gender ratio of medical doctors in Sweden have changed in the last seven years, as discussed in the background section. According to SCB (the Central Bureau of Statistics in Sweden), female medical doctors have gone from 13 700 in practitioners in 2005 to 16 400 in 2012, while in the same time period male practitioners have declined from 18 700 to 18 000.

This trend of equalisation between male and female medical doctors might have had an effect on the Swedish population. Granted that being a medical doctor infers intelligence and competence in and of itself, and the availability of male and female doctors equalise, gender as a tool of discriminating between competence and intelligence will decrease and people will start using other cues to assess these qualities. This also includes, one can argue, femininity and masculinity as ways of inferring competence and intelligence, since they are bound to gender categorisation.

What remains then is the mismatch effect discussed previously. If classical gender based stereotypes are less common when assessing medical doctors, then a mismatch effect explains why female characters with mixed traits received lower average ratings for competence and intelligence.

However, as femininity and masculinity seem to play a smaller role in inferring competence and intelligence, emphatic abilities are still associated with femininity for the female characters. This might be explained by a weaker association between empathy and the medical doctor profession, which then do not infer with the predicted stereotype. Instead femininity might still be more widely used as a cue of empathic abilities when assessing medical doctors.

However, to draw anything more than these speculative conclusions, further investigation is acquired. A recommended next step is to incorporate a wider array of stimuli. A larger set of different voices that are controlled for more characteristics than fundamental frequency, and perhaps real life people as visual stimuli. Also, a larger sample of participants that are more geographically diverse is also suggested. This setup may be used to investigate and compare gender based stereotypes, or lack thereof, for both characters that are medical doctors and for characters that exhibit no profession, to see how far the results in this study can be generalised.

Finally, if these results reflect a real way of thinking and inferring intelligence, competence, and empathic abilities that can be generalised to a broader population, it tells us something of the inner workings of stereotyping, which in turn can be used to predict attitude responses to societal changes. Furthermore it can be used as an argument for affirmative action. If theses findings is corroborated by future research, indeed changing gender distributions in certain domains may help to eliminate otherwise common stereotypes, where gender or other arbitrary attributes has been used to infer other personal qualities.

References

Cadinu, M., Maass, A., Rosabianca, A. and Kiesner, J. (2005). Why Do Women Underperform Under Stereotype Threat? Evidence for the Role of Negative Thinking. *Psychological Science*, 16, 572-578.

Blair, I., Steiner, J., Faircloughe, D., Hanratty, R., Price, D., Hirsch, H., Wright, L., Bonsert, M., Karimkhani, E., Magid, D. and Havranek, E. (2013). Clinicians'Implicit Ethnic/Race Bias and Perceptions of Care Among Black and Latino Patients. *Annals of Family Medicine*, 11, 43-52.

Gigerenzer, G. & Goldstein, D. (1996). *Reasoning the Fast and Frugal Way: Models of Bounded Rationality*. Psychological Review, 103, 650-669.

Gigerenzer, G., Todd, P. and ABC Research Group (1999). *Simple Heuristics That Make Us Smart*. New York: Oxford University Press.

Goldstein, D. and Gigerenzer, G. (2002). Models of Ecological Rationality: The Recognition Heuristic. *Psychological Review*, 109, 75-90.

Gulz, A., Ahlner, F. and Haake, M. (2007). Visual femininity and masculinity in synthetic characters & patterns of affect. In Paiva, A., Prada, R. & Picard, R. (eds.), *Affective Computing and Intelligent Interaction* (ACII) (pp. 654-665). Lisbon, Portugal, 12-14 September.

Haake, M. & Gulz, A. (2008). Visual Stereotypes and Virtual Pedagogical Agents. *Educational Technology & Society*, 11, 1-15.

Kahneman, D. (2011). *Thinking, Fast and Slow*. London: Penguin Books Ltd.

Hilton, J., & von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology,* 47, 237-271.

Latu, I., Stewart, T., Myers, A., Lisco, C., Estes, S. and Donahue, D. (2011). What We "Say" and What We "Think" About Female Managers: Explicit Versus Implicit Associations of Women With Success. *Psychology of Women Quarterly*, 35, 252-266.

Nass, C. and Brave, S. (2005). *Wired for Speech – How Voice Activates and Advances the Human-Computer Relationship*. Cambridge: The MIT Press.

Reeves, B. and Nass, C. (1996). *The Media Equation – How People Treat Computers, Television, and New Media Like Real People and Places*. Stanford: CSLI Publications.

Sabin, J., Marini, M. and Nosek, B. (2012). Implicit and Explicit Anti-Fat Bias among a Large Sample of Medical Doctors by BMI, Race/Ethnicity and Gender. *PloS ONE*, 7, 1-7.

SCB (Statistiska Centralbyrån) (2013). *Genomsnittlig grund- och månadslön samt kvinnors lön i procent av mäns lön efter region, sektor, yrke (SSYK) och kön. År 2005 – 2012*, [Electronic]. Available at: http://www.scb.se/Pages/ProductTables.aspx?id=14374

Steffens, M. and Jelenec, P. (2011). Separating Implicit Gender Stereotypes regarding Math and Language: Implicit Ability Stereotypes are Self-serving for Boys and Men, but not for Girls and Women. *Sex Roles*, 64, 324-335. PLOS

Tversky, A. and Kahneman, D. (1986). Rational Choice and the Framing of Decisions. *The Journal of Business*, 59, 251-278.

Tversky, A. and Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 4, 207-232.

Williamson, G. (2006). *Human communication: a linguistic introduction* (2 ed.). Billingham: Speech-Language Services.