Weiss, A., & Bartneck, C. (2015). Meta Analysis Of The Usage Of The Godspeed Questionnaire Series. Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN2015), Kobe pp. 381 - 388. | DOI: 10.1109/ROMAN.2015.7333568

Meta Analysis Of The Usage Of The Godspeed Questionnaire Series

Astrid Weiss[1] and Christoph Bartneck[2]

Abstract— **Standardized metrics for assessing the success of robots is a necessity for a research field to compare and validate results. The Godspeed Questionnaire Series (GQS) is one of the most frequently used questionnaires in the field of Human-Robot Interaction (HRI) with over 160 citations as of October 2014. In this paper, we present a meta analysis of studies that used the GQS. The HRI community uses a large variety of robotic platforms and only the NAO robot seems to be used by multiple research groups. A qualitative meta analysis of 18 NAO studies reveals accumulated findings on perceived intelligence, likability, and anthropomorphism, but also reveals contradictions on how the robot's behaviour and task context impact GQS ratings. The paper closes with a reflection on how added value of data analysis and presentation could be achieved for the HRI community in future.**

I. Introduction

Measurement and replicability are still prominent topics in Human-Robot Interaction (HRI) research. To develop as a research field HRI findings obtained by one research group need to be reproduced by other groups. However, in most cases comparability can hardly be achieved due to different user study methods, different robots, and different tasks the user has to perform with the robot. Already in 2006 Steinfeld et al. [1] called for standardized metrics that would be compiled into an HRI Metric Toolkit. Such a toolkit "allows for greater sharing of knowledge as it becomes possible to compare findings and to benchmark designs". Steinfeld et al. approached the question predominantly from a technical perspective, but also included a social component to assess the quality of the interaction from a user perspective. During the last five years one measurement tool became popular in the HRI community to evaluate the perception of social interactions with robots: The Godspeed Questionnaire Series (GQS) [2]. According to Google Scholar, it is the most highly cited paper in the International Journal of Social Robotics and with currently 160 citations it is cited more frequently than any paper published at the ACM/IEEE International Conference on Human-Robot Interaction so far. This does not mean that the GQS is necessarily the best measurement tool, it only means that it has been widely cited.

Due to the fact that this questionnaire is open access, simple to use, and available in several languages, we now have the unique opportunity to perform a secondary analysis, meaning comparing which results have been achieved with this questionnaire so far and interpreting them from a meta perspective. This will enable the HRI community to gain knowledge on which overall results could be achieved with this measurement tool until now. Such a meta perspective is not unusual as Deth pointed out: "The idea that collecting your own data is the ideal situation for researchers is based on a clear misunderstanding of the role of empirical testing and exploration in research and also on an overestimation of the need for newly collected data. In fact, using existing data is the rule rather than the exception in social research." [3]. Unfortunately, due to the above mentioned problems, such a data analysis is difficult in the young discipline of HRI. Until now, only very few studies were ever replicated and barely any standardized metrics have been established.

The GQS does allow us to attempt a first secondary analysis of published data on user assessments of HRI scenarios. According to Hyman a secondary analysis of survey data is defined as: "the extraction of knowledge on topics other than those which were the focus of the original survey" [4]. In our case we are interested in gaining insights in two main aspects (1) How did other researchers refer to the GQS - How do they use it, cite it and (2) What findings can be accumulated with the GQS so far - Does the comparison of the results of all these studies allow us to gain insights on its applicability and which independent variables affect it?

For instance, the appearance of robots certainly affects their perception as has been shown in empirical studies (e.g. [5]), so it could be assumed that the GQS could be rated very similar for all studies involving the same robot, even though it is used in very different scenarios. However, also the interaction with the robot and its displayed behavior during this interaction can be assumed to have a higher impact than the mere appearance and can change the perception of the robot.

In this regard, it is crucial to understand by which factors the measurement of the GQS is influenced on the side of the user (e.g. personality, background, pre-experience), on the side of the robot (e.g. appearance, verbal and nonverbal behavior) and with regard to the specific scenario both interaction partners are placed in, which can be determined amongst others by the task context and the application scenario. Therefore our analysis was guided by the following leading questions:

1) To what robots was the GQS applied?
2) How often were different robots compared with each other?

We greatly acknowledge the financial support from the Austrian Science Foundation (FWF) under grant agreement T623-N23, V4HRC.

[1] Astrid Weiss is with the Institute of Automation and Control, Vienna University of Technology, Gusshausstrasse 27-29/E376, 1040 Vienna, Austria astrid.weiss@tuwien.ac.at

[2] Christoph Bartneck is with the HIT Lab NZ, University of Canterbury, Private Bag 4800, 8140 Christchurch, New Zealand christoph.bartneck@canterbury.ac.nz

3) What behaviors of the robot(s) were manipulated?
4) In what context did the interaction take place?
5) What application fields did the researchers have in mind for the robot(s)?
6) How often did the researchers use the Wizard-of-Oz technique?
7) How many participants were used in the experiments?
8) In what country was the study conducted?

II. The Godspeed Questionnaire Series - GQS

The GQS consists of five scales that are relevant to evaluate the perception of (social) Human-Robot Interaction. The scales are *Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety*. The GQS is available in English, German, Spanish, Chinese, Japanese, French, Greek, Arabic, and Dutch. The scales consist of five-point semantic differentials such as "Fake — Natural". The GQS is available for free on the internet[1]. There is a slight overlap between the Anthropomorphism and Animacy Scales, since the item "Artificial — Lifelike" appears in both.

Ho and MacDorman were concerned that the Godspeed indices may not measure the concept after which it was named, but instead measure some convolution of that concept and interpersonal warmth [6]. They then proceeded with proposing an alternative to the GQS in the form of the four factors: attractiveness, eeriness, humanness, and warmth. One would have to doubt whether the data gathered by Ho and MacDorman can be the basis of any evaluation of the GQS, since only video clips of computer animated characters and robots were used. They did not use the stimuli for which the GQS was intended: actual robots. Nevertheless, it has to be acknowledged that the measurement tools themselves are being tested by replicating studies. And no matter if there may or may not be a better way to measure the concepts of the GQS, for our approach to analyse and synthesise research literature on the GQS it is only relevant that a large number of studies have used the same questionnaire. Ho and MacDorman's study has currently be only cited 42 times according to Google Scholar and this is much below the 160 citations that the GQS has received.

Ho and MacDorman were not the only ones to develop questionnaires that intend to measure the impression that users have of robots. Kamide et al. published a series of papers in which they proposed new measurement scales [7]. Nomura et al. proposed several scales to measure the negative attitude towards robots [8] and anxiety towards robots [9]. Caine et al. [10] developed the Collected Robot Scale which is a combination of the GQS with the 6-level scale for "fear", "suprise", "disgust", and "unpleasantness" proposed by Nonaka et al. [11]. Similarly, Moshkina et al. [12] used the GQS as starting point to develop eight novel semantic differential scales, which were tested in two live HRI experiments with a NAO robot. Again, none of these scales have been as widely used as the GQS.

An interesting study was performed by Ruijten et al. [13] who designed 37 yes/no items about anthropomorphic qualities a robot can have. After validating this scale they evaluated in how far it relates to other measurements of anthropomorphism and compared it to the Waytz scale [14] and the anthropomorphism scale of the GQS. The results suggested that all three scales measure the same concept, which is already a positive indication for the applicability of the GQS for HRI studies. However, the aim of our meta analysis presented in this paper is not a statistical validation or the summary of findings, but to yield new conclusions by pulling together fragmented results of single studies using the GQS.

III. Method

Our analysis was a non-statistical meta analysis following a similar approach as described in [15]. We searched for all the papers that cite the GQS paper [2] in Google Scholar (GS). GS has the widest coverage from all the scientific literature search engines and is therefore most suitable for our analysis [16], [17], [18]. Other search engines, such as Scopus or the ACM Digital Library offer more structured and reliable data at the cost of being less inclusive. There is a very strong correlation between the ACM Digital Library and GS, but GS is collecting four times as many citations in the HRI discipline [19]. We downloaded all the papers that were available to us or our colleagues and imported the PDF files together with the citation information into an Endnote Database. This database was the foundation for our meta analysis.

After a first analysis of the 160 papers it became obvious that the research presented in these manuscripts was not amenable to the methods of statistical meta-analysis due to well-known limitations in HRI research, such as small sample sizes, variety of different platforms etc. Therefore, a non-statistical analysis and synthesis was required. Coding articles is used in such an analysis in order to be consistent, replicable, and valid. After reading all the articles, only papers that did not actually mention the GQS in the text and duplicates were eliminated by hand. Next, we identified in the coding procedure which papers actually used the GQS in some way in an empirical study.

The further analysis of the empirical work was then guided by one main assumption: Depending on the application context/scenario, the experimental design, the cultural context, and the independent and dependent variables in experimental designs, results of the GQS might differ, e.g. a stationary NAO robot will be potentially perceived as safer than a moving one. Such assumptions are needed in a non-statistical meta analysis in order to move the body of knowledge forward to explain a bigger phenomenon: "Like the blind man's description of the elephant, [...] primary studies may provide many conclusions, but little understanding of the big picture" [15]. After screening all empirical work again we decided to analyse all papers involving research with NAO robots for a qualitative interpretations as those provided the best foundation with respect to our previous assumptions. In

[1] http://www.bartneck.de/2008/03/11/the-godspeed-questionnaire-series/

the following we will present the results of each analysis step.

IV. Results

From the 160 citations we found, 23 were of books or book chapters, 55 were journal articles, 57 were conference papers, 15 were theses, 9 were reports, and one was an unpublished work. The papers were published since the year 2009 and Table I shows the number of papers per year that referenced the GQS. We were unable to retrieve the actual papers from seven of the references, since they were either listed in Google Scholar only as a citation, they appeared in books to which we had no access or no further trace of the reference could be found outside of Google Scholar.

Year	# papers
2009	11
2010	13
2011	27
2012	24
2013	52
2014	33
Total	**160**

TABLE I

Papers per year

In total, 21 papers had an overlap with at least one other paper and to avoid double counting we excluded 9 papers. The overlap typically occurred when authors published the results of the same study in different publication venues. It might have first appeared as a conference paper before it was then improved into a journal article. Some papers, in particular the reports, occasionally reported on more than one study and hence it was necessary to break up such a report into several papers. The final number of studies to be included in our analysis was therefore 144.

Overall, 75 papers referenced the GQS without using it in an empirical study, which brings down the number of available studies to 69. It needs to be mentioned that we also excluded papers at this point, which used the GQS as basis to develop a different questionnaire, which was the case for 9 studies. In total 35 different robots were used for these 69 studies which indicates a high fragmentation of the robots used by the HRI community. 62 studies only used one robot, four used two robots and two studies used more than two robots. Of the 69 studies, the only robotic hardware that was used frequently by more than one research group was the NAO. 18 studies used this robot for an empirical study in which the GQS was used as a dependent variable. The GQS consists of five scales and their usage frequency in all 69 studies is presented in Table II.

Scale	Count
Anthropomorphism	38
Animacy	38
Likeability	46
Perceived Intelligence	46
Perceived Safety	37

TABLE II

Count of GQS scales

Originally, we intended to compare studies with similar set-ups which report the mean values and standard deviations of the GQS and perform statistical calculations based on these results. However, only 43 publications mentioned these statistics. Others only reported differences which were found or not found. The studies used such a broad variety of robot types, different independent variables, and interaction scenarios that such a quantitative comparison would have been too unstable for a valid statistical comparison. The number of observations for each specific condition would have been too low for the effect sizes we typically expect in HRI experiments. Instead we decided to focus on the NAO studies as they offer one stable factor, namely the appearance of the robot and therefore allow us a to interpret the findings achieved with the GQS in a qualitative manner.

V. NAO Studies

When focusing on the 18 NAO studies, we notice in an overview (see Table III) that only six studies used all scales of the GQS, a slight preference can be seen for measuring perceived intelligence and likeability.

Ten of the studies were carried out in the Netherlands, three in Germany, two in New Zealand, and one in Denmark, Egypt & Japan, and the USA. Unfortunately, an identification of the questionnaire language used in the studies was very difficult as only two of the papers reported the language used; it can be assumed that the questionnaires were used in the native language of the participants, however, it was not directly mentioned.

In all studies just one NAO robot was used, except for [20] which was picture-based and used in total 33 different picture sets including one with NAO depicted. In four studies NAO acted completely autonomously, in seven studies it was mentioned that the robot was pre-programmed and sometimes the otherwise autonomous interaction was initiated remotely. Five papers clearly mentioned that the robot was controlled using the Wizard-of-Oz technique. However, this categorization did not apply to study [20], since it used a picture stimulus. From the 13 studies in which the robot was pre-programmed or wizarded, 11 used the perceived intelligence scale of the GQS as dependent variable.

The envisaged application context was mentioned in eight of the 18 studies and was in six cases elderly care, however, it has to be mentioned that these six studies were conducted in the framework of the same EU-project which focused on the development of a smart home environment with an integrated robot to enable independent aging in place.

With respect to the interaction scenarios, dialogue and communication situations most often offered the basis for a study. Direct contact interaction with the robot was rather the exception (see [21] and [22]) .

A. Perceived Intelligence

[23] revealed that perceived intelligence is more affected by the interaction scenario than the actual behaviour of the

id	Application Context/Scenario	Mode	Scales	Country	N	UV
[23]	elderly care/approach directions	WoZ	all	NL	14	3 different approach behaviours
[24]	no context/casual conversation (led by robot)	Auton.	PI, Ant	NL	60	no conditions
[25]	no context/persuasive story-telling	Auton.	all	NL	48	different gaze and gestures combinations
[26]	no context/robot as cleaner or guide	Auton.	PI	NL	45	introvert/extrovert robot as cleaner or guide
[27]	robot bartender/social engag. w.customers	Auton.	PI, LI,	GER	48	before/after interaction
[28]	Care/robot as care-giver	Auton.	PI	USA	60	robot as doctor vs. robot as patient
[29]	no context / conversation	WoZ	PI	DK	41	nodding vs. not nodding
[22]	no context/human-like motion copying	Auton.	Ant.	EG/JP	36	different imitation conditions
[30]	elderly care /conversation	Auton.	all, not Ant.	NL	19	2 gaze behaviours: looking-while-talking vs. looking-while-listening
[21]	no context/turn taking	Auton.	all	GER	28	four different behaviours (exploration, interaction, avoid interaction, full interaction)
[31]	social assistance/solving mental rotation tasks	Auton.	all	GER	20	two interaction strategies: structuring interaction and performance-based interaction
[32, p.80]	elderly care/robot in the smarthome	WoZ	LI, Ani.	NL	10	NAO vs. Smarthome for message delivery
[32, p.82]	elderly care/robot as exercise coach	WoZ	LI	NL	16	before/after
[32, p.84]	elderly care/robot in the smart home	Auton.	LI	NL	16	NAO vs. smarthome
[33]	no context/message retention	Auton.	all	NL	23	before/after interaction; NAO experience yes/no
[34]	elderly care/ robot attracting attention	WoZ	all	NL	12	4 different behaviours for attracting attention (eye-contact, blinking eyes, gesture, saying hello)
[20]	no context/ anthropomorphisation	pic.-based	Ant.	NZ	51	upright /inversed pictures
[35]	no context/play quiz with robot	WoZ	PI	NZ	40	emotionality/no emotionality; intelligence/no intelligence

TABLE III

Overview on NAO Studies: WoZ=Wizard of OZ, Auton.=Autonomous; PI=Perceived Intelligence, LI=Likeability, Ant.=Anthropomorphism, Ani.=Animacy

.

robot. This goes in line with the finding from [24] that the perceived realism of an interaction with a robot can serve as a predictor and explain perceived intelligence ratings. [27] even found a decrease in perceived intelligence ratings after the interaction and argue that the limited task domain (NAO serving as a bar tender) could have caused the low rating.

However, the findings of [26] are in contradiction to this. In this study which especially focused on how the task context and the personality of the robot impact the GQS ratings no significant differences could be found for any of the scales. [28] in comparison, could demonstrate that the role of the robot impacts the perceived intelligence ratings: NAO in the role of a doctor was perceived as more intelligent than NAO in the role of a patient.

Findings from [29] and [21] support the assumption that more complex behaviour fosters higher ratings on the perceived intelligence scale, which was also demonstrated in a well-designed complex experiment with the iCub robot, in which the robot performed six scenarios which differed in interaction complexity [36].

An interesting finding regarding perceived intelligence was finally retrieved in [20], namely that participants with pre-experience in interacting with NAO rated its perceived intelligence significantly lower than participants without any pre-experience.

However, details were missing on what pre-experience with NAO exactly meant: Were participants aware of how to programme NAO or did they take part in other studies with NAO before. In the second case it would be relevant to

Weiss, A., & Bartneck, C. (2015). Meta Analysis Of The Usage Of The Godspeed Questionnaire Series. Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN2015), Kobe pp. 381 - 388. | DOI: 10.1109/ROMAN.2015.7333568

know more about the previous studies in order to put it into context and interpret it correctly.

B. Likability

[23] revealed that it is not the style of the approaching behaviour, but the type of interaction scenario which causes higher ratings on the likability scale. This finding again supports the assumption that the task context is more important than the interaction behaviour. The importance of the task context on likability ratings is also visible in [32], where NAO was compared to a smart home environment to communicate messages. NAO was rated as more likable which could also be interpreted as more suitable for the task context. This result was also confirmed in a follow-up field trial [32]) where NAO was rated again as more likable than the smart home environment.

C. Anthropomorphism

[24] revealed that high anthropomorphism ratings can be an indicator for social acceptability of the NAO robot and [35] that higher perceived intelligence ratings are not an indicator for higher anthropomorphism ratings. [22] demonstrated the importance of the interaction situation in relation to anthropomorphism assessments. The authors could show that even a simple manipulation of having the participant imitating the robot for a few minutes positively impacts the later assessment of the human-likeness of the robot's motion. [20] could demonstrate with the help of the anthropomorphism scale that the inversion effect (the phenomenon that images of human faces and bodies are harder to recognize upside down than pictures of objects) also applies to images of robots which are similar to humans. Unfortunately, the paper only reveals that besides pictures of other robots also NAO pictures were used as stimulus, but it was not stated in the conclusion if the NAO robot was categorized as "similar to humans", but only that all anthropomorphic robot stimuli were categorized like that.

D. Robot Behaviour and Task Context

As mentioned before it stays unclear looking at the results of the 18 NAO studies in how far the robot behaviour and the task context affect the ratings of the GQS. For instance, [25] concludes that their study does not provide evidence for different effects of gazing or gestures on any of the GQS dimension, but [30] concludes that when the robot reacts to the gaze behaviour of a person, better ratings on all the GQS scales can be assumed. In [31] the authors conclude (although they do not explicitly relate their conclusion to GQS results) that the subjective ratings of the robot indicate that a performance-based interaction strategy (i.e. the robot reacts situation-specific to the user) can help improving user performance on difficult cognitive tasks. Interestingly, only two studies summarized that they could not find any significant differences in the GQS ratings [26], [34] with respect to robot behaviour and task context.

In [26] the authors assumed that the appropriate personality for a robot depends on the task context and compared an introvert-cleaning NAO to an introvert-tour guide, to an extrovert-cleaning NAO, and an extrovert-tour guide. They found trends that preferences for robot personalities may be dependent on the task context, however, they did not find a significant interaction effect between the robot's behaviours and the task context for the perceived intelligence scale. In [34] the authors wanted to explore which action is best suited for (re)gaining a person's attention and implemented four different behaviours on NAO: (1) NAO says "Hello", (2) NAO attempts to make eye-contact, (3) NAO flashes the LED lights which represent its eyes, (4) NAO waves to the participant. However, no significant differences for any of the GQS scales could be found.

E. Before and After Comparisons

In total nine studies performed a before/after comparison using the GQS. However, it is hardly ever mentioned what participants actually ranked in the "before" condition. Did they see the robot, did it actually do something? Before/after comparisons are very typical for instance for the NARS questionnaire [8], however, in this case an attitude change is measured and participants will always have an attitude about robots also before seeing one and as other studies showed it often changes to a more positive attitude after the interaction [37] (an effect that might be caused by the novelty effect). However, against the assumption of fellow researchers the GQS is not intended "to provide indications of people's attitudes toward robots" [38], it is intended to assess a robot in terms of specific quality criteria, which hardly can be done without experiencing the robot interaction.

One interesting idea was found in [33] where the authors intended to use the GQS as manipulation check to see if the novelty effect influences the ratings: "If the interactions had a large influence on the results of the Godspeed questionnaire, this could indicate that the robot presented a high level of novelty for subjects". However, the conclusion that no differences were found before and after the interaction does not necessarily need to mean that there was no novelty effect, but that the interaction was not complex enough to see significant changes in the interaction as indicated by the work of [36], [21].

F. Reflection on the NAO Studies

Summarizing the accumulated main findings from the 18 NAO studies we can retrieve interesting insights on three of the five scales. Regarding the perceived intelligence scale, the interaction scenario, the complexity of the behaviour, the role of the robot, and the realism of the interaction seem to be relevant impact factors. For likability the interaction scenario and the appropriateness of the application context seem to play a role. Regarding anthropomorphism the findings suggest that even small manipulations such as movement imitation or inversion can impact anthropomorphism ratings and that high ratings on anthropomorphism are a predictor for high ratings on social acceptance.

However, the question in how far the behaviour of the robot and the task context affect the assessment of the GQS cannot be satisfyingly answered by the means of our meta analysis. Contradicting results were found in the papers which cannot be resolved purely by using the presented information. More observations are necessary to test this specific aspect. However, our analysis reveals open research topics which could be closed by future controlled studies using the GQS. It would deepen our understanding of the measurement tool itself on the one hand and on the other hand it would enhance the understanding of the impact of robot behaviour and task context on subjective experience ratings.

VI. Discussion

One of the biggest challenges for a meta analysis in earlier days was to identify which research to include and how to judge its validity (e.g. distinguishing grey literature from published literature). This became easier through the Internet and the data gathering revealed to us the benefit of Open Access publishing. Often we were not able to access a publication from the publisher, since our universities did not have the specific subscription. Even amongst four researchers from four different universities in four different countries it was not possible to acquire all publications. Only because authors and research organizations posting their publications in their open repositories it was possible to collect the majority of the publications. This highlights the benefits of making scientific literature open to the public. In particular meta studies like the one at hand benefit from openly available literature, but also every other researcher who wants to survey the state of the art needs to have access to the latest results.

The results from our analysis showed that almost all studies surveyed only used one robot in their experiment. Barely any study compared different embodiments. 62 studies only used one robot and 35 different robots were used. Only the NAO robot was used by multiple studies in multiple research organizations. This fragmentation of robot hardware use makes it very difficult to compare results. The reason for this fragmentation might be that research groups cannot afford multiple robots or that many labs build their own robots. The GQS is then used to evaluate their own designs. A huge variety of robots used in HRI can also be seen as an advantage. Engineers are constantly trying to build and program better robots. However, this also makes comparisons difficult, in particular if only a very small number of robots are being produced. The latest development of open source hardware might show the way towards a more collaborative approach on building robots. The necessary files to 3D print an InMoov robot, for example, can be freely downloaded [39]. Similarly, the documentation for the electronic components, such as the micro controllers and servo motors, are openly available. Sharing the development of a robot might lead to robots being more widely available. Building a robot using 3D printing and standard micro controllers, such as Arduino boards, make the production of a robot also much cheaper than proprietary designs.

At first sight it is surprising that 13 studies attempted to measure the perceived intelligence of the NAO robot despite the fact that a human operator actually controlled the robot. It can be argued that the participants rated the intelligence of the operator, not the one of the robot. Often the main reason for using a Wizard-of-Oz methodology is the poor speech recognition quality of the NAO robot [40]. The operator then usually follows a strict interaction script by only selecting the next action sequence for the robot.

A. Methodology

The results showed that all five scales within the GQS have been used frequently. No particular scale was used noticeably more often than another one. This indicates that all the concepts that the GQS tries to measure appear to be relevant to HRI researchers for (social) interaction evaluation.

We observed that there have been nine studies that applied the GQS before and after the participants interacted with a robot. While such a procedure might be appropriate to test the potential change in attitudes towards robots in general, it remains doubtful whether applying the GQS prior to interacting with robots results in any useful data. The GQS is not intended to test attitudes, but to test the direct impression that users have of a specific robot.

What we also noted in several publications was that the semantic differential was often also called 5-point Likert scale which is wrongly used terminology. A Likert scale consists of several items (questions or statements) which have to be rated with a score from a pre-defined range (eg. 1-5). A semantic differential in comparison asks respondents to choose between to opposite poles, offering a pre-defined range with eg. five steps. Another important remark at this point is that in several studies the GQS was used with 7-point or 10-point scales instead of 5-point. Actually, changing the answer format would request checking the internal and external validity of the scale again [41]. Moreover, changing the answering scale makes it more difficult to directly compare study results.

Furthermore, only in one publication we found a clearly mentioned downside of using the GQS in an empirical study: “This matches the comments of most of the participants of our experiments which complained about the similarity between many questions and about some high-level attributes which were difficult to assign to the robot” [21]. Remarks like that are needed to further develop measurement tools, but are often omitted in scientific publications.

VII. Conclusions

From the 160 citations, only very few were useful for a meta analysis. This clearly indicates that our community is still dominated by research groups that develop their own robotic hardware and measurements instruments. The call of Steinfeld et al. [1] for more standardized metrics, at least for the measuring the social aspects of HRI does not yet have come to life.

A question Hyman already posed for social science surveys can by now also be posed for user-centered HRI studies: “Why has relatively little scientific wealth been extracted from the figurative mountain of gold?” [4]. For a long time one could have argued in HRI, we are lacking measurement tools or comparable platforms, however despite the enormous increase in the amount of user study data generated, the potential of using existing data is evidently still far from being exhausted. An incredible amount of self-reporting and behavior observation data has been stored in the last ten years and is waiting to be reanalyzed by creative and skeptical researchers to replicate or verify assumptions and build a theoretical grounding for HRI. For future HRI research it would be beneficial to set up collaborative projects addressing the need to harmonize data and to develop further standardized instruments such as the GQS as that would imply an expansion of research opportunities. Furthermore, the increasing body of data means that comparative analysis over for example cultures, robot embodiments, and usage scenarios could become reality if the community starts not only sharing code but also user study data. This would offer opportunities researchers of an earlier era could only dream of. Our research community is aware on how to use modern communication and information technologies, we would just need to make our data openly accessible.

To conclude, our strategy for conducting a non-statistical meta analysis allowed us to explore our assumptions on how the application/context/scenario and the experimental design affect the scales of the GQS. Moreover, it gave us insights on research procedures in the HRI community an enabled us to make suggestions for improvements in study designs and results reporting to enable better (and maybe even statistical) meta analysis in the future.

VIII. Limitations

The sample of literature we used for our analysis is not representative for the whole HRI literature since we only selected papers that referenced the GQS. Other measurement tools might have received less citations, but could have potentially be used more often in empirical studies. We intend to investigate other popular measurement tools for HRI in the future. Another approach to prove the validity of self-reporting results would be the systematic combination of the GQS with an implicit/behavioural measurement, as it has already be performed once by fellow researchers for the perceived safety scale [42].

References

[1] A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. Goodrich, “Common metrics for human-robot interaction,” in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*, 2006, pp. 33–40.

[2] C. Bartneck, E. Croft, D. Kulic, and S. Zoghbi, “Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots,” *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009.

[3] J. van Deth, “Using published survey data,” in *Cross-cultural survey methods*, J. A. Harkness, F. J. Van de Vijver, and P. P. Mohler, Eds. Wiley-Interscience Hoboken, 2003, vol. 325.

[4] H. Hyman, *Secondary Analysis of Sample Surveys: Principles, Procedures, and Potentialities*. Wiley: New York, 1972.

[5] A. M. Rosenthal-von der Pütten and N. C. Krämer, “How design characteristics of robots determine evaluation and uncanny valley related responses,” *Computers in Human Behavior*, vol. 36, pp. 422–439, 2014.

[6] C.-C. Ho and K. F. MacDorman, “Revisiting the uncanny valley theory: Developing and validating an alternative to the godspeed indices,” *Computers in Human Behavior*, vol. 26, no. 6, pp. 1508 – 1518, 2010.

[7] H. Kamide, Y. Mae, K. Kawabe, S. Shigemi, M. Hirose, and T. Arai, “New measurement of psychological safety for humanoid,” in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, 2012, pp. 49–56.

[8] T. Nomura, T. Kanda, and T. Suzuki, “Experimental investigation into influence of negative attitudes toward robots on human-robot interaction,” *AI & Society*, vol. 20, no. 2, pp. 138–150, 2006.

[9] T. Nomura, T. Suzuki, T. Kanda, and K. Kato, “Measurement of anxiety toward robots,” in *The 15th IEEE International Symposium on Robot and Human Interactive Communication, ROMAN2006*. IEEE, 2006, pp. 372–377.

[10] K. Caine, S. Sabanovic, and M. Carter, “The effect of monitoring by cameras and robots on the privacy enhancing behaviors of older adults,” in *7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2012, pp. 343–350.

[11] S. Nonaka, K. Inoue, T. Arai, and Y. Mae, “Evaluation of human sense of security for coexisting robots using virtual reality. 1st report: evaluation of pick and place motion of humanoid robots,” in *Robotics and Automation, 2004. Proceedings. ICRA’04. 2004 IEEE International Conference on*, vol. 3. IEEE, 2004, pp. 2770–2775.

[12] L. Moshkina, “Reusable semantic differential scales for measuring social response to robots,” in *Proceedings of the Workshop on Performance Metrics for Intelligent Systems*. ACM, 2012, pp. 89–94.

[13] P. A. Ruijten, D. H. Bouten, D. C. Rouschop, J. Ham, and C. J. Midden, “Introducing a rasch-type anthropomorphism scale,” in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. 2559825: ACM, 2014, pp. 280–281.

[14] A. Waytz, C. K. Morewedge, N. Epley, G. Monteleone, J.-H. Gao, and J. T. Cacioppo, “Making sense by making sentient: effectance motivation increases anthropomorphism.” *Journal of personality and social psychology*, vol. 99, no. 3, p. 410, 2010.

[15] C. J. Bland, L. N. Meurer, and G. Maldonado, “A systematic approach to conducting a non-statistical meta-analysis of research literature.” *Academic Medicine*, vol. 70, no. 7, pp. 642–53, 1995.

[16] L. I. Meho and K. Yang, “Impact of data sources on citation counts and rankings of lis faculty: Web of science versus scopus and google scholar,” *Journal of the American Society for Information Science and Technology*, vol. 58, no. 13, pp. 2105–2125, 2007.

[17] L. I. Meho and Y. Rogers, “Citation counting, citation ranking, and h-index of human-computer interaction researchers: A comparison of scopus and web of science,” *Journal of the American Society for Information Science and Technology*, vol. 59, no. 11, pp. 1711–1726, 2008.

[18] K. Kousha and M. Thelwall, “Sources of google scholar citations outside the science citation index: A comparison between four science disciplines,” *Scientometrics*, vol. 74, no. 2, pp. 273–294, 2008.

[19] C. Bartneck, “The end of the beginning - a reflection on the first five years of the hri conference,” *Scientometrics*, vol. 86, no. 2, pp. 487–504, 2011.

[20] J. Zlotowski and C. Bartneck, “The inversion effect in hri: are robots perceived more like humans or objects?” in *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*. IEEE, 2013, pp. 365–372.

[21] G. Schillaci, S. Bodiroza, and V. V. Hafner, “Evaluating the effect of saliency detection and attention manipulation in human-robot interaction,” *International Journal of Social Robotics*, vol. 5, no. 1, pp. 139–152, 2013.

[22] Y. Mohammad and T. Nishida, “Human-like motion of a humanoid in a shadowing task,” in *International Conference on Collaboration Technologies and Systems (CTS)*. IEEE, 2014, pp. 123–130.

[23] M. T. Bruna, “The benefits of using high-level goal information for robot navigation,” Masters Thesis, Eindhoven University of Technology, 2011.

[24] M. M. A. de Graaf and S. Ben Allouch, “Exploring influencing variables for the acceptance of social robots,” *Robotics and Autonomous Systems*, vol. 61, no. 12, pp. 1476–1486, 2013.

[25] J. Ham, R. Bokhorst, R. Cuijpers, D. van der Pol, and J.-J. Cabibihan, *Making Robots Persuasive: The Influence of Combining Persuasive Strategies (Gazing and Gestures) by a Storytelling Robot on Its Persuasive Power*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, vol. 7072, ch. 8, pp. 71–83.

[26] M. Joosse, M. Lohse, J. G. Perez, and V. Evers, “What you do is who you are: The role of task context in perceived social robot personality,” in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2013, pp. 2134–2139.

[27] S. Keizer, P. Kastoris, M. E. Foster, A. Deshmukh, and O. Lemon, “Evaluating a social multi-user interaction model using a nao robot,” in *ROMAN*. IEEE, 2014, in press.

[28] K. J. Kim, E. Park, and S. Shyam Sundar, “Caregiving role in humanrobot interaction: A study of the mediating effects of perceived benefit and social presence,” *Computers in Human Behavior*, vol. 29, no. 4, pp. 1799–1806, 2013.

[29] A. Krogsager, N. Segato, and M. Rehm, *Backchannel Head Nods in Danish First Meeting Encounters with a Humanoid Robot: The Role of Physical Embodiment*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2014, vol. 8511, ch. 62, pp. 651–662.

[30] S. Patelski, “The effect of gaze behavior on the attitude towards humanoid robots,” Bachelor Thesis, Eindhoven University of Technology, 2012. [Online]. Available: http://home.ieis.tue.nl/etorta/reports/BEP_Patelski.pdf

[31] S. Schneider, I. Berger, N. Riether, S. Wrede, and B. Wrede, *Effects of Different Robot Interaction Strategies During Cognitive Tasks*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7621, ch. 50, pp. 496–505.

[32] E. Torta, J. Oberzaucher, F. Werner, R. H. Cuijpers, and J. F. Juola, “Attitudes towards socially assistive robots in intelligent homes: results from laboratory studies and field trials,” *Journal of Human-Robot Interaction*, vol. 1, no. 2, pp. 76–99, 2012, 17.

[33] E. T. van Dijk, E. Torta, and R. H. Cuijpers, “Effects of eye contact and iconic gestures on message retention in human-robot interaction,” *International Journal of Social Robotics*, vol. 5, no. 4, pp. 491–501, 2013.

[34] J. van Heumen, “Human robot interaction: Acquiring attention,” Tech. Rep., 2014. [Online]. Available: http://home.ieis.tue.nl/rcuijper/reports/BEP_JimvanHeumen_verbeteringen4.pdf

[35] J. Zlotowski, E. Strasser, and C. Bartneck, “Dimensions of anthropomorphism: from humanness to humanlikeness,” in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. 2559679: ACM, 2014, pp. 66–73.

[36] V. Vouloutsi, K. Grechuta, S. Lallee, and P. F. Verschure, *The Influence of Behavioral Complexity on Robot Perception*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2014, vol. 8608, ch. 29, pp. 332–343.

[37] N. Mirnig, E. Strasser, A. Weiss, and M. Tscheligi, “Studies in public places as a means to positively influence people’s attitude towards robots,” in *Social Robotics*, ser. Lecture Notes in Computer Science, S. Ge, O. Khatib, J.-J. Cabibihan, R. Simmons, and M.-A. Williams, Eds. Springer Berlin Heidelberg, 2012, vol. 7621, pp. 209–218.

[38] E. v. Dijk, “Investigating rejection behavior in the ultimatum game as a measure of anthropomorphism,” 2013.

[39] G. Langevin, “InMoov Project,” 2014. [Online]. Available: http://www.inmoov.fr/project/

[40] O. Mubin, J. Henderson, and C. Bartneck, “You just do not understand me! speech recognition in human robot interaction,” in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2014, (in press).

[41] N. Schwarz, B. Knäuper, H.-J. Hippler, E. Noelle-Neumann, and L. Clark, “Rating scales numeric values may change the meaning of scale labels,” *Public Opinion Quarterly*, vol. 55, no. 4, pp. 570–582, 1991.

[42] S. Zoghbi, E. Croft, D. Kulic, and M. Van der Loos, “Evaluation of affective state estimations using an on-line reporting device during human-robot interactions,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2009, pp. 3742–3749.