HAL
open science

Effectiveness and usability of technology-based interventions for children and adolescents with ASD: A systematic review of reliability, consistency, generalization and durability related to the effects of intervention

Cécile Mazon, Charles Fage, Hélène Sauzéon

To cite this version:

Cécile Mazon, Charles Fage, Hélène Sauzéon. Effectiveness and usability of technology-based interventions for children and adolescents with ASD: A systematic review of reliability, consistency, generalization and durability related to the effects of intervention. Computers in Human Behavior, 2019, 93, 10.1016/j.chb.2018.12.001 . hal-01950078

HAL Id: hal-01950078

https://inria.hal.science/hal-01950078v1

Submitted on 10 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Effectiveness and Usability of Technology-based Interventions for Children and Adolescents with ASD:

A Systematic Review of Reliability, Consistency, Generalization and Durability Related to the Effects of Intervention

---

Mazon, Cécile[1,2] ; Fage, Charles[3] ; Sauzéon, Hélène[2,4]

[1] Equipe-projet Phoenix, INRIA Centre de Recherche Bordeaux Sud-Ouest – 200 Avenue de la Vieille Tour, 33405 TALENCE Cedex, Gironde, Aquitaine, France

[2] Laboratoire EA 4136 Handicap, Activité, Cognition, Santé, Université de Bordeaux – escalier 1B Batiment Laboratoires site Carreire, 146 rue Léo Saignat, 33076 BORDEAUX Cedex, Gironde, Aquitaine, France.

[3] Unité de logopédie clinique, Faculté de Psychologie et des Sciences de l'Éducation, Université de Liège (Belgique) – Bâtiment B33, Boulevard du Rectorat, 3, 4000 Liège (Saint Tilman), Belgique.

[4] Equipe-projet Flowers, INRIA Centre de Recherche Bordeaux Sud-Ouest – 200 Avenue de la Vieille Tour, 33405 TALENCE Cedex, Gironde, Aquitaine, France

---

Correspondence concerning this article should be addressed to Cécile Mazon, Equipe-projet Phoenix, INRIA Bordeaux Sud-Ouest, 200 Avenue de la Vieille Tour, 33405 TALENCE Cedex, Gironde, Aquitaine, France.

E-mail: cecile.mazon@inria.fr

Compliance with Ethical Standards

Conflicts of interest: The authors declare that they have no conflict of interests

Research involving Human Participants and/or Animals: This article does not contain any studies with human participants or animals performed by any of the authors.

Acknowledgement. This research was possible with the support of French Orange Foundation and INRIA Bordeaux Sud-Ouest research center. The authors thank *Aquitaine Traduction services* for the final revision of the manuscript.

Highlights

- We reviewed technology-based interventions (TBI) for children and youths with ASD
- We separated Therapeutic Effectiveness (TE) and Technology Usability (TU) studies
- TE studies were more compliant with methodological standards than TU studies
- Studies exploring TE and TU emerged as promising interdisciplinary approaches
- Both study design and measure reliability affected the strength of evidence for TBI

Effectiveness and Usability of Technology-based Interventions for Children and Adolescents with ASD: A Systematic Review of Reliability, Consistency, Generalization and Durability Related to the Effects of Intervention.

Abstract. A growing number of studies have investigated technology-based interventions (computer, phone, tablet, robot, etc.) for supporting children and teenagers with ASD, notably in school settings. Past reviews stressed study-design weaknesses of TBI researches. This systematic review has threefold purpose: 1) to update the previous ones with a focus on clinical-quality studies; 2) to examine reliability, consistency, durability and generalization of measurements; and 3) to compare the methodology of two cores of studies according to two dimensions: *Therapeutic Effectiveness (TE)* and *Technology Usability (TU)*. From the 685 search results, 31 studies were selected (22 on TE, 6 on TU, and 3 on TE-TU). Overall, few studies reached the standards of evidence-based practices (reliability, consistency, durability, generalization). TE studies provided more evidence of their reliability than TU and TU-TE studies. Moreover, the examination of studies' results revealed that: 1) the more robust study designs, the less consistent TBI effect, 2) the more reliable the measure, the less large TBI-related effect size. Although less robust, TE-TU studies can be seen as an emerging interdisciplinary approach, combining expertise in human-computer interaction and clinical research.

Keywords. *Autism spectrum disorder, Technology-based interventions, Therapeutic effectiveness, Usability, Systematic review, Methodology.*

Introduction

Autism spectrum disorder (ASD) refers to a neurodevelopmental disorder with two main characterized symptoms, varying in severity across the spectrum: 1) impaired communication and social interactions, and 2) restricted activities and interests (such as repetitive behaviors and stereotypies) (American Psychological Association [APA], 2013). From a very young age, ASD affects the entire range of daily living activities, restricting social participation of individuals. As a result, they struggle with difficulties to be enrolled at school, or to find and keep a job (*e.g.*, Reed & Osborne, 2014; Taylor, Henninger, & Mailick, 2015). To address this situation, a growing number of studies in recent decades have explored the opportunities for technology-based intervention (TBI) for supporting children and teenagers in their daily life, notably in school settings. Technologies such as computer-based tools, virtual/augmented reality, mobile- and tablet-based applications, as well as robotics, are now considered promising approaches for designing interventions for ASD, targeting various outcomes, such as social and academic skills, on-task and challenging behaviors, *etc.* (*e.g.*, Begum, Serna & Yanco, 2016; Grynszpan, Weiss, Perez-Diaz, & Gal, 2014). As individuals with ASD are keen on using digital devices, this avenue of research has been receiving a lot of attention (Odom, et al., 2015), with studies examining the feasibility and the effectiveness of TBIs. The purpose of this systematic review is to evaluate current research in TBI to promote school-related capabilities in children and adolescents with ASD. Specifically, to move forward the field, it focuses on the studies' validation methodologies, by screening design and outcome measurements for both therapeutic effectiveness and usability of the technologies involved.

Previous reviews: main findings and limitations

Several literature reviews have been published about the use of technologies in interventions with children and adolescents with ASD (*e.g.*, Grynszpan, et al., 2014; Ploog, Sharf, Nelson, & Brooks, 2013; Knight, McKissick & Saunders, 2013; Odom, et al., 2015). Each review stressed specific findings regarding "sub-fields" of interest in ASD TBIs. First, when considering the evidence according to the type of technology (*e.g.*, robotic, Begum, Serna & Yanco, 2016; computers, Ploog, et al., 2013; Ramdoss, Lang, et al., 2011; Ramdoss, Mulloy, et al., 2011; Ramdoss, et al., 2012), computer-based interventions have apparently attracted numerous controlled studies aimed at proving TE (*e.g.*, Ploog, et al., 2013), while more recent technologies, such as robotics, have received less attention (Begum, Serna & Yanco, 2016). Second, as the range of outcomes in ASD interventions is wide, some reviews have focused on certain types of processes or behaviors (*e.g.*, academic skills, Knight, McKissick & Saunders, 2013; communication, Ramdoss, Lang, et al., 2011; literacy skills, Ramdoss, Mulloy, et al., 2011; social and emotional skills, Ramdoss, et al., 2012). These studies drew some positive conclusions

concerning the efficacy of TBIs for a variety of target skills, but the strength of evidence is again limited, due to poor-quality study design, indicating that TBI are, at best, promising/emerging practices (Knight, McKissick & Saunders, 2013; Ramdoss, et al., 2012). Finally, other reviews investigating the age range in TBI studies pointed out that they mainly target preschool- and school-aged children with ASD. Yet, considering the poor outcomes in adulthood with ASD, adolescents have considerable needs for intervention, especially towards the end of compulsory education and during the transition to adulthood (Odom, et al., 2015).

A lot of these reviews pinpointed design weaknesses in the studies claiming to provide evidence for the efficacy of technology-based interventions: the study design is often reported as too weak, due to small sample sizes, or even the absence of a comparative control group. As a result, these weaknesses are advanced as a main explanation of the evidence inconsistency. Previous reviews did not systematically use objective scales for design assessment, even for systematic review purposes (*e.g.*, Ploog, et al., 2013; Ramdoss, Lang, et al., 2011; Ramdoss, et al., 2012). The strength of study design may be quantified with specific rating scales, formalizing and hierarchizing the levels of evidence according to acknowledged methodological criteria (*e.g.*, Scottish Intercollegiate Guidelines Network [SIGN], 2008; Jadad, et al., 1996). Basically, the gold standard in clinical trials is the randomized controlled trial (RCT), with a very stringent study design, in order to provide the hardest evidence of therapeutic effectiveness and minimize the risk of bias. RCTs involve at least two comparable groups, with random allocation unknown to both experimenters and participants. Compared to RCTs, controlled experimental trials with a pre-post design are less stringent, but provide pilot assessments of the effects of an intervention.

Additionally, none of these reviews addressed the value of outcomes' measurements, particularly with regards to the distinctions between standardized *vs.* non-standardized measurements, and between objective *vs.* subjective measurements. For this latter distinction, although subjective measurements are useful for screening people's feelings and opinions, they are subjects to several bias, such as social desirability, self-assessment reliability or inter-rater reliability (Annett, 2002). At the other hand, objective measurements may be more reliable since they often rely on performance or factual observations. Even if both still are complementary for fully evaluating the effects of intervention, objective measurements may provide a higher level of evidence than subjective ones. Whatever the measurement is objective or subjective, the consistency and the reliability of evidence may be improved by using standardized measurements. Among clinical outcomes, the measurements from standardized clinical tests are recognized to be reliable, while non-standardized measurements, such as those obtained in *ad hoc* tests, are less acknowledged, due to lack of evidence of their validity and reliability,

which increases the risk of a false measurement (Drost, 2011). Moreover, the use of standardized tests allows comparing and replicating studies with reliable and consistent outcome measurements.

The aim of standardized measurements is a selective investigation of the integrity of each cognitive process or behavior, in order to identify specific cognitive or behavioral deficiencies associated with pathologies. Two categories of standardized clinical tests may be distinguished: formal *vs.* naturalistic (Chan, Shum, Toulopoulou, & Chen, 2008). For instance, by asking participants to name the emotion depicted on each face, Ekman's facial emotion recognition test (Ekman & Friesen, 1976) assesses the cognitive process of recognizing facial emotions; while the Social Responsiveness Scale (SRS, Constantino & Gruber, 2005) assesses social capabilities through items related to everyday situations. Correlations between these two kinds of measurements are often not significant, as they result from separate constructs (process *vs.* activity) and/or self-rating biases evoked for naturalistic tests based on subjective measurements (Toplak, West & Stanovich, 2013). Both formal and naturalistic standardized tests contribute together to the ASD diagnosis (Taylor, et al., 2016; Volkmar, et al., 2014), by addressing the overall functioning of individuals with ASD (Chan, et al., 2008).

The distinction between formal *vs.* naturalistic tests raises the question of the ecological value of outcome measurements. The ecological validity (*i.e.*, the extent to which an outcome measurement is similar to real-life activities) provides evidence for the transfer of the intervention's effects to everyday life. Ecological validity is measured on two criteria (Kenworthy, Yerys, Anthony, & Wallace, 2008): the extent to which the measurement correlates with an individual's everyday performance (*veridicality*), and/or the extent to which the measurement mirrors the demands of the everyday environment (*verisimilitude*), as provided by the naturalistic tests mentioned above. An intervention demonstrates strong evidence of generalization when the study shows positive effects on everyday-like tasks linked to its outcome (ecological transfer).

Regardless of measurement reliability (standardized *vs.* non-standardized measurements) and ecological value, the consistency and the durability of outcome measurements are also expected when assessing the effect of an intervention. Durability refers to the length of time therapeutic effects are maintained (Ardoin, 2006) and is typically assessed with a short- to long-term follow-up, to distinguish the near and far effects of the intervention. Consistency is assessed by examining internal and external validity (Simms, 2008). Internal validity refers to measuring the target process or behavior to provide evidence in favor of the intervention. External validity refers to measuring other processes or behaviors to ensure that the intervention has no effects other than the target outcome. In other words, an intervention exhibits strong evidence of TE when positive effects on the target outcome are observed, but no other effects (particularly negative) on other cognitive processes and behaviors.

To sum up, previous reviews gave an insight into various technologies, target processes and behaviors, as well as age range, and clearly documented the weaknesses in study design (*e.g.*, Ploog, et al., 2013; Ramdoss, Lang, et al., 2011; Ramdoss, Mulloy, et al., 2011; Ramdoss, et al., 2012). For this reason, the purpose of this review is to update and enrich the previous reviews, with a focus on the studies with the most robust study designs. As the quality of intervention measurements is a critical requirement for evidence-based practice (Grondin & Schieman, 2011), the review of reliability, ecological value, as well as consistency and durability of TBIs in ASD studies may provide new insights for understanding their actual effects.

Two distinct purposes: TE vs. TU?

To the best of our knowledge, none of previous reviews addressed the ergonomic issue of usability of the technology. Studies examining the effects of TBIs have primarily explored the therapeutic effects of such interventions and put themselves in the field of health interventions assessment. Therapeutic effectiveness (TE) refers to the extent to which an intervention improves a relevant clinical outcome (*e.g.*, skill, behavior*, etc.*) for the studied population. This concept is closely related to the field of clinical studies, and to the requirements of evidence-based practices for evaluating the effects of interventions. Providing evidence for the therapeutic effects of an intervention is of primary importance for validating the use of TBIs as remediation and support tools with individuals with ASD.

However, another point of equal importance is to address the issue of prerequisite skills for benefiting from a given TBI. Numerous studies in Human-Computer interaction have described interface requirements suited to the specific needs (notably perceptual and sensory-motor skills) of individuals with ASD interacting with technology (*e.g.*, Hayes, et al., 2010; Hourcade, Williams, Miller, Huebner, & Liang, 2013; Putnam & Chong, 2008). This issue may be explored by ergonomic observation of the usability and accessibility of the intervention technology (Hersh, 2014; Inostroza, Rusu, Roncagliolo, & Rusu, 2013). According to the International Organization for Standardization (ISO), usability refers to *"the extent to which a product, a system or a service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"* (ISO/IEC 9241-11, Bevan, Carter & Harker, 2015). Accessibility is defined as "*the usability of a product, a system or a service, environment or facility by people with the widest range of capabilities"* (ISO, 2014: ISO/IEC Guide71). The ISO 9241-11 definition of usability identifies three key dimensions of usability: (1) *effectiveness*, the extent to which the task is appropriately completed by the user; (2) *efficiency*, the ability to reach the specified goal with minimum resources; and (3) *satisfaction*, the willingness to use the product and comfort level when using it (Bevan, Carter & Harker, 2015). If the product is

not usable, the user will misuse or disregard the product, or even abandon its use (*i.e.*, because the product makes it impossible to complete the task, is too inefficient, or is uncomfortable for the user). Consequently, TU deserves an in-depth examination in the field of TBI for individuals with ASD, as the TU is the vehicle of intervention, *i.e.*, the key to accessing its content. TU should be a precondition for any TBI investigation since it may positively or negatively impact the magnitude of the intervention effect. TU also acts as part of the experimental control by guaranteeing the proper administration of the intervention to the participant.

Basically, TU evaluations are both objective and subjective: the former focuses on the effectiveness and the efficiency, while the latter mostly concerns the user satisfaction. Objective measurements of usability are quantitative performance data, often derived from technology use scenarios, such as the success rate and time required to complete the task (Baharuddin, Singh, & Razali, 2013). Subjective evaluations are obtained via user interviews and questionnaires. There are emergent ways to assess objectively the user satisfaction by using physiological measurements (*e.g.*, electrodermal response, gaze patterns) probing the emotional responses during the use of the system (Agarwal & Meyer, 2009; Sharafi, Soh, & Guéhéneuc, 2015). Standardized TU measurements have also been developed, such as the System Usability Scale (SUS, Brooke, 1996) and the Quebec User Evaluation of Satisfaction with assistive Technology (QUEST, Demers, Weiss-Lambrou, & Ska, 2000). An overview of usability engineering methods can be found in Holzinger (2005).

Aim and contributions

Our general purpose was to review the data from ASD studies, using TBIs to enhance cognitive processes and/or school-related capabilities (*e.g.*, academic skills, such as literacy or calculation, adaptive behaviors, such as autonomy, social interactions, and communication). Specifically, we reviewed studies focusing on TE and/or TU.

This review aimed to make a twofold contribution for advancing the state-of-the-art in the field of TBIs for ASD. First, an in-depth examination of measurement quality in the more robust studies was conducted, according to specific rating scales (Jadad, et al., 1996; SIGN, 2008). Outcome measurements were analyzed in terms of reliability (standardized measurements or not), consistency (internal and external measurements in relation to therapeutic target), ecological value (generalization or transfer) and durability (near/far effects). Second, the distinction between two purposes of studies (*i.e.*, TE and TU) may provide new insights in the research practices aiming at evaluating or validating the use of TBI with ASD population.

Method

Search procedure

A systematic literature search was conducted in online databases linked to the scientific fields relevant to both technologies and ASD interventions: *PubMed, IEEE Xplore, ACM Digital Library, Springer, Taylor & Francis, Scopus, Education Resources Information Center (ERIC), ScienceDirect/Elsevier and EBSCO (PsycArticles, PsychInfo, Psychology and Behavioral Sciences Collection).* The selection was limited to peer-reviewed articles in English, published between January 2000 and September 2016. The search query was built using keywords linked to our topic, according to the PICO criteria (Table 1). [Insert Table 1]. After a first screening for duplicates and non-English references, titles, abstracts and keywords were examined to exclude irrelevant articles. Finally, we iteratively applied inclusion and exclusion criteria to the remaining references. We also verified that multiple articles published by the same research group did not contain overlapping data. When this was the case, we retained the article with the most comprehensive information.

Selection procedure

Each article was reviewed by the author and coded for inclusion and exclusion criteria. Any doubts were resolved with a second evaluator. The following inclusion criteria were used: (a) the study involved a TBI/training, (b) participants included children or adolescents (0-20 years-old) formally diagnosed with ASD, (c) it evaluated the TE and/or TU of a TBI, (d) it addressed assistance with and/or remediation of cognitive processes and/or school-related skills (*e.g.* communication, socialization, engagement behavior *vs.* hand-washing, cooking), and (e) the design was sufficiently robust (assessed by SIGN and Jadad ratings, see Appendix).

The following were excluded: (a) articles that did not report the study method and/or results, (b) research based on a single- or multiple single-case design, (c) the study did not address an issue relating to the technology itself (*e.g.*, learning procedure or behavioral training to use the technology), (d) the technology evaluated was not designed for use by a child (*e.g.*, teleconferencing to train parents/professionals in intervention techniques, data analysis support for therapists, *etc.*) and (e) the technology used was not interactive. Indeed, some techniques, such as video modeling, do not require actions by the child: the interaction is passive, while other interventions used interactive supports, requiring active participation by the child. Like Grynszpan, et al. (2014), we excluded this type of TBI, in order to focus on technologies actively used by the child. In the same way, we excluded studies where the technology was not used by the children themselves, to focus on devices suitable for use by a child with ASD.

Data extraction and categorization

The remaining references were checked using the SIGN ratings for levels of evidence (SIGN, 2008), to select the most robust studies for this review (see details in the Appendix). SIGN ratings were used to exclude poorly-designed studies, such as non-comparative and non-controlled pre-post designs. To reduce the number of references for this review, we only included studies rated 1++, 1+, and 1- (randomized and non-randomized controlled trials). A Jadad score, ranging from 0 to 5 points, represented a quick, easy tool for an additional assessment of methodological quality (Jadad, et al., 1996). Studies with a score above 2 were considered high quality, whereas a minimum score of 2 was acceptable if it was not possible for the design to be double-blind.

The remaining articles were screened to extract the following data: authorship, year of publication, TE and/or TU study, group characteristics (N, Age, medical condition), technology used, aim of intervention, study settings, intervention duration, research design, outcomes measurements, and results.

The effect sizes for TBI outcomes were computed from the results reported within each study, using Microsoft Excel software (version 16.14). The effect sizes of outcome measures were then averaged for each study. When it was possible, Cohen's *d* were computed from means and standard deviations when both were numerically reported in the study. Otherwise, Cohen's *d* were computed, if applicable, from either eta-squared or test statistics (t-test or one-way ANOVA). We were unable to compute effect sizes for three studies, due to the lack of data in the reporting (Fage, 2015; Jeong, et al., 2015; Valadão, et al., 2016). Formulas for computing Cohen's *d* were retrieved from Ellis (2010), McCartney & Rosenthal (2000) and Fritz, Morris, & Richler (2012).

Results

Literature search and quality ratings

[Insert Figure 1]

Among the 917 references extracted from database search results, we identified 232 duplicates and non-English papers. The 685 remaining references were then checked for inclusion on the basis of title, abstract, and keywords, there were 283 potential papers for inclusion. A further 204 references were eliminated on the basis of our inclusion and exclusion criteria. The SIGN ratings of levels of evidence (SIGN, 2008) were applied to the 79 remaining articles to exclude studies of insufficient quality: 31 papers scored 1, no studies scored 2, and 48 studies scored 3. Only the articles that scored 1 were included, leading to consider 31 studies. Figure 1 depicts the flow diagram for selecting the articles. The Jadad scale (Jadad, et al., 1996) was applied to the remaining 31 papers: only one was rated 3, 7 were rated 2, 12 were rated 1, and 11 were rated 0. None of the studies reported a double-blind design. Results will be described according to study's characteristics as following: 1) general description of deployed TBIs; 2) participants and study design; 3) outcomes measurements; 4) results and effect sizes. Main information on included studies are presented according to the study purpose: TE studies in Table 2, TU studies in Table 3 and TE-TU studies in Table 4. [Insert Table 2, 3 & 4].

Studies' purposes classification and general description of interventions (technology, target outcomes, settings)

Thirty-one studies met the inclusion criteria and were included in this review. Out of these 31 studies, 3 addressed both TE and TU (TE-TU studies), 6 focused on TU assessment, and 22 focused on evaluating TE. The results are described below and then compared according to the issues addressed by the study (*i.e.*, TE or/and TU).

A majority of reviewed TBIs involved computer- and robot-based interventions. Computer-based interventions often consisted in game software designed for enhancing facial expression and emotion recognition: *FaceMaze* (Gordon, Pierce, Bartlett, & Tanaka, 2014), *FaceSay* (Hopkins, et al., 2011; Rice, Wall, Fogel, & Shic, 2015), *EmotionTrainer* (Silver & Oakes, 2001). The computer game software *TeachTown* addressed a broader set of skills, including social, as well as cognitive and academic skills (Whalen, et al., 2010). These four programs were evaluated in terms of TE with children with ASD (3-13 years old). Unless for *FaceMaze*, which was assessed in a single session at the laboratory (Gordon, et al., 2014), the evaluation of these TBIs took place at school over a period ranging from 2 weeks to 3 months (Hopkins, et al., 2011; Silver &

Oakes, 2001; Rice, et al., 2015). Two further TE studies were conducted on computer programs for addressing communication skills in school-aged children with ASD. Grossman, Peskin & San Juan (2013) designed the *Gruffee task* for training children to communicate about their actions and evaluated the communicative clarity after a week of training at the laboratory. Ploog, Banerjee & Brooks (2009) evaluated prosody comprehension in a single session with a computer game involving a cartoon bird searching for nuts, which triggered different spoken sentences. Among TU studies, a further computer-based intervention involved three games for enhancing socio-emotional skills: *What to choose?, Intruder* and *Faces*. Participants used the game *What to choose?* over a period of 3 months at school, while the two other games were dedicated to the display evaluation (Grynszpan, Martin & Nadel, 2008). Zheng, Warren, et al. (2016) conducted a single-session laboratory study for evaluating the TU of a computer-based learning environment in an early social orienting training for toddlers with ASD. One last computer-based intervention with a tangible interface addressed only pre-academic skills (shape and color recognition) in preschoolers with ASD. Sitdhisanguan, Chotikakamthorn, Dechaboon, & Out (2012) evaluated both TE and TU in an overtime clinic with two separate session: after one week of use in their first evaluation, and after four weeks in the second.

Robot-based interventions mainly consisted in robot-mediated training for enhancing either emotional, social and/or communication skills. Interestingly, the series of TE studies conducted by Srinivasan, Park, Neelly, & Bhat (2015), Srinivasan, Eigsti, Gifford, et al. (2016), and Srinivasan, Eigsti, Neely, et al. (2016) reported results of the same 8-weeks rhythmic intervention based on the robot Nao. Each study reported results about different outcomes: social (Srinivasan, Eigsti, Neely, et al., 2016) and communication skills (Srinivasan, Eigsti, Gifford, et al., 2016), as well as emotional skills and repetitive behaviors (Srinivasan, et al., 2015). The robot Nao was used in two further TBIs designed for preschoolers with ASD (Bekele, Crittendon, Swanson, Sarkar, & Warren, 2014; Zheng, Young, et al., 2016). In their TE study, Zheng, Young, et al. (2016) evaluated imitation skills after a single-session training involving Nao. Bekele, et al. (2014) evaluated the TU of their Nao-based system for enhancing joint attention abilities. A further TU study was conducted with the robot *MARIA* for improving social skills in children with ASD (Valadão, et al., 2016). These two TU studies were also conducted in a laboratory with a single session assessment. Five remaining TE studies involved robot-based interventions. In two studies, Pop, et al. (2013) and Pop, Pintea, Vanderbroght, & David (2014) used the robot *Probo* for enhancing social skills (Pop, et al., 2013, 2014), as well as play and engagement skills (Pop, et al., 2014). Salvador, Silver & Mahoor (2015) designed an intervention with the robot *Zeno* for improving emotion recognition. Costescu, Vanderborght & David (2015) evaluated the TE of an intervention based on the robot *Keepon* in a reverse learning task for enhancing cognitive flexibility. These four robot-based interventions were

all evaluated for their TE in a single session. Finally, Jeong, et al. (2015) evaluated the TE of using the robot *iRobi Q* for enhancing emotional vocabulary after 20 sessions with a frequency of 1-2 per week.

Other kind of technologies were assessed for TE purposes. Golan, et al. (2010) and Young & Posselt (2012) evaluated the video DVD *The Transporters* for enhancing emotional and social skills. Both studies took place at home, for a period of 3 and 4 weeks. The four remaining TE interventions were based on 1) virtual-reality for enhancing emotional skills (Lorenzo, Lledó, Pomares, & Roig, 2016), 2) Kinect motion-based games for enhancing attention and visuo-motor skills (Bartoli, Garzotto, Gelsomini, Oliveto, & Valoriani, 2014), 3) multitouch tabletop for enhancing social skills (Bauminger-Zviely, Eden, Zancanaro, Weiss, & Gal, 2013), and 4) a tablet-based application for enhancing language skills (Rodriguez & Cummings, 2016). Among TU studies, Bekele, et al. (2013) evaluated a virtual environment for training facial expression in adolescents with ASD, and Falkmer, et al. (2014) assessed a smartphone-based system for supporting autonomous school transportation in children with disabilities. Both studies implemented their evaluation in a single session. Two last studies evaluated both TE and TU of tablet-based applications designed for supporting emotion regulation (Fage, 2015) and the realization of school activities (Fage, Pommereau. Consel, Balland, & Sauzéon, 2016). These studies took place at school for a 3-month period.

To sum up, the most frequently evaluated technologies were computer- (N= 10) and robot-based interventions (N= 11). Also, robot-based interventions were more often evaluated in a single session (N= 7/11), whereas computer-based ones were frequently evaluated after at least one week of use (N= 7/10). Social, emotional and/or communication skills were the primary target outcomes (N= 23/31), which is in line with ASD-related impairments. TU studies implemented more single session at the laboratory (N= 5/6), while TE studies involved more longitudinal evaluation (N= 14/22), and even some ecological settings (N= 10/22). In the same line, TE-TU studies involved an evaluation period from one week to three months, and their settings were quite ecological: two evaluations were conducted at school (Fage, 2015; Fage, et al., 2016). The last one was conducted in an overtime clinic, where participants were used to receiving their treatment (Sitdhisanguan, et al., 2012). This result is not surprising since TU evaluations are often conducted after a single use of the system, through scenarios, performance measures and/or questionnaires. Conversely, TE studies need a minimal intervention period for allowing the TBI to elicit substantial benefits that can be captured by the measurements. By contrast, ecological settings should deserve more consideration for both TE and TU purposes, keeping in mind that the controlled environment provided by the laboratory compromises the chances to catch real-life outcomes.

Study designs' screening (participants, inclusion/exclusion, design)

The 31 studies included represent a total of 796 participants. Importantly, the three studies conducted by Srinivasan, et al. (2015), Srinivasan, Eigsti, Gifford, et al. (2016), and Srinivasan, Eigsti, Neelly, et al. (2016) reported different results from the same sample of 36 children with ASD. Twenty-one studies involved school-aged children (range 5-12 years), five studies involved adolescents (range 13-18) and five involved preschoolers (0-5 years). Of the 21 studies that reported gender distribution (N= 515 participants), 424 participants were male and 91 female, *i.e.,* 82% male participants.

Two studies recruited children with disabilities (Falkmer, et al., 2014; Rodriguez & Cummings, 2016), but did not report the distribution of ASD *vs.* other disorders (*i.e.*, Down Syndrome and Speech-Language Impairment). In the remaining studies (N = 728 total participants), a total of 576 participants had an ASD diagnosis (*approx.* 79%). According to the distinction between low- and high-functioning ASD (LF-ASD and HF-ASD) depending on the co-occurrence of an intellectual deficiency (IQ ≤ 70), 17 studies reported the level of functioning of their participants: 10 studies recruited participants with HF-ASD, 6 recruited participants with LF-ASD, and one recruited both HF-ASD and LF-ASD participants.

Because HF- or LF-ASD conditions as well as ASD severity may influence TBI outcomes, it is critical to screen participants' characteristics with reliable tools. Thus, we reviewed the use of standardized tests for recruitment and inclusion/exclusion purposes.

TE studies. Out of the 22 TE studies, 13 studies used standardized clinical tests as inclusion/exclusion criteria, including 9 that used ASD diagnosis scales (Bauminger-Zviely, et al., 2013; Costescu, Vanderborght, & David, 2015; Golan, et al., 2010; Gordon, et al., 2014; Grossman, Peskin & San Juan, 2013; Hopkins, et al., 2011; Pop, et al., 2014; Whalen, et al., 2010; Zheng, Young, et al., 2016). Participants' cognitive functioning was controlled in terms of intellectual functioning and/or verbal abilities in 10 studies (Bauminger-Zviely, et al., 2013; Golan, et al., 2010; Gordon, et al., 2014; Grossman, Peskin & San Juan, 2013; Hopkins, et al., 2011; Jeong, et al., 2015; Pop, et al., 2014; Rice, et al., 2015; Silver & Oakes, 2001; Young & Posselt, 2012). Social impairment was controlled in 3 studies (Bauminger-Zviely, et al., 2013; Young & Posselt, 2012; Zheng, Young, et al., 2016). Grossman, Peskin & San Juan (2013) also included visual perception and motor coordination, as well as Theory-of-Mind (ToM) measurements in their recruitment procedure, assessed using the Beery VMI developmental test (Beery & Beery, 2004) and the ToMi (Hutchins, Prelock & Bonazinga, 2012), respectively.

TU studies. Standardized clinical tests were used in three studies as inclusion/exclusion criteria for medical conditions, as well as group matching (Bekele, et al., 2013, 2014; Grynszpan, Martin & Nadel, 2008). The most widely-used scales concerned ASD diagnosis, intellectual functioning, and social abilities: for

instance, the ADOS (Lord, et al., 2000), SRS (Constantino & Gruber, 2005), SCQ (Rutter, Bailey & Lord, 2003), and WASI (Wechsler, 2014) were used by Bekele, et al. (2013; 2014) to assess participants formally diagnosed with ASD as well as TD participants; the WISC (Wechsler, 2003) was used by Grynszpan, Martin & Nadel (2008) to verify the intellectual functioning of participants with ASD.

TE-TU studies. Two studies used standardized tests as inclusion/exclusion criteria for the ASD diagnosis or for group matching on intellectual functioning (Fage, 2015; Fage, et al., 2016). Social impairment in natural settings was also assessed in one study (Fage, et al., 2016) using the SRS (Constantino & Gruber, 2005).

In summary, studies examining TE and/or TU of TBIs mainly targeted school-aged children (N= 21/31). TE studies used more often standardized clinical tests for depicting participants' characteristics before recruitment. Of the 6 TU studies, only two strictly verified the ASD condition, using a standard ASD diagnosis scale. Conversely, among the 22 TE studies, 13 verified the ASD diagnosis using standardized scales and gave clinical details on their samples. Some studies reported minimal data about participants, asking for the replicability of their protocol. The use of standardized measurements for recruitment procedures have often concerns with either the confirmation of ASD diagnosis and their intellectual abilities. However, they rarely took account of ASD-related specificities such as ASD symptoms severity or their particularisms in perceptual style or motor skills, as done by Grossman, Peskin & San Juan (2013). Yet, ASD specificities provide relevant information for recommending a TBI with respect to the needs and abilities of individuals with ASD. More than validating TBI for individuals with ASD, a relevant survey of participants' characteristics may allow recommending TBI with respect to individuals' needs and abilities. ASD is characterized by a large heterogeneity across individuals and TBI may have differential effects depending on users' characteristics (*e.g.,* cognitive functioning, motor skills).

Let us now have a look on the study designs and the sample sizes, as well as the reporting of drop-outs. For TBI studies, dropouts may inform on eventual usability or acceptability problems with the technology.

TE studies. Sample sizes across TE studies ranged from 5 to 41 participants per group, with an average around 15 participants per group. According to the Jadad scale, a majority of the TE studies scored between 1 and 3 (only 3 studies scored 0). This set included 13 studies that reported excluded/dropped out participants (N= 38; 9% on average). The most frequent reason for dropping out was refusal or no interest in 5 studies (N= 15 participants). Other reasons for dropping out were: incomplete data in 2 studies (N= 3), abandonment in one study (N= 1), excessively severe impairments in 3 studies (N= 4), and moving or hospitalization in 2 studies (N= 4). Surprisingly, 11 participants were excluded from one study due to unusable data (Gordon, et al., 2014): participants were filmed during facial emotion production but excluded when the facial emotion was not

sufficiently visible. Thirteen TE studies were RCTs, with random group allocation among participants with ASD. Also, one TE study can be qualified as quasi-RCT (Golan, et al., 2010), involving two randomly-allocated groups with ASD participants, as well as a control group with typically-developed (TD) participants. A further TE study adopted a group-based crossover design, where the treatment and control groups were switched in the middle of the intervention (Bauminger-Zviely, et al., 2013). The 7 remaining TE studies were all controlled trials, including 3 studies involving only participants with ASD, and the 4 others involving participants with ASD and typically-developed ones. Among TE studies, 18 used a pre-post design, while 3 others compared the target intervention with another type of intervention (Costescu, Vanderborght & David, 2015; Pop, et al., 2013; Zheng, Young, et al., 2016), and one simply compared participants with and without ASD (Salvador, Silver & Mahoor, 2015).

TU studies. Sample sizes across studies ranged from 5 to 23 participants per group, with an average around 10 participants per group. According to the Jadad scale, 5 of out the 6 studies scored 0 and the remaining study scored 1, thanks to the inclusion of a statement about dropouts (6 participants were excluded due to refusal or distress; Bekele, et al., 2014). All TU studies were controlled trials, involving a treatment group composed with ASD participants, and a control group with typically-developed ones. Only one study had a pre-post design (Grynszpan, Martin & Nadel, 2008), and all studies manipulated two factors: medical conditions (*e.g.*, ASD *vs.* TD) and/or several intervention conditions (*e.g.*, robot *vs.* human; rich *vs.* simple interfaces).

TE-TU studies. Sample sizes across studies ranged from 4 to 8 participants per group, with an average around 5 participants per group. All three TE-TU studies scored 0 on the Jadad scale since none of them were RCTs and documented any dropouts. The three TE-TU studies were all controlled trials but unlike TU studies, they had all a pre-post design. Also, one study compared three intervention conditions (Mouse *vs.* WIMP *vs.* tangible interface; Sitdhisanguan, et al., 2012). Two studies recruited only participants with ASD and non-randomly allocated them to conditions (Sitdhisanguan, et al., 2012; Fage, 2015). The last study recruited participants with ASD as the treatment group, and participants with ID as the control group (Fage, et al., 2016).

To sum up, sample sizes were larger in TE studies than in TU and TE-TU studies. On the whole set of studies, 79% of participants were individuals with ASD but all TU studies recruited typically-developed participants as control group. Conversely, 16 TE studies recruited only participants with ASD and only two TE studies directly compared performances of ASD vs. typically-developed participants. For the remaining TE studies, control groups with typically-developed participants were dedicated to contrasting pre-post differences in the treatment group. The majority of TE studies implemented an RCT and/or employed a pre-post design, while TU studies were all controlled trials with only one adopting a pre-post design. All the TE-TU studies were

controlled trials with a pre-post design, unlike TU studies. Drop-out were reported in only a half of the set of included studies, including 13 with a TE purpose. This result must be seen in relative terms since some studies may not deplore drop-outs during their evaluation, and then did not report their absence. However, it remains surprising that only one TU study reported this information because dropouts may inform about technology acceptability, which is related to TU.

Measurements' screening

After screening study designs' characteristics across studies, we focused on the measurements used for assessing interventions' outcomes, with respect to four dimensions: reliability, consistency, durability, and generalization.

Evaluating the effects of TBI: reliability and consistency of measurements

Consistency. Regarding internal and external validity of intervention studies (*i.e.* consistency), all the studies reviewed assessed the direct outcomes of TBI (internal validity), but not the side effects in the extra-domains of the TBI target (external validity). However, the analysis of side effects is as important as that of the direct outcomes, particularly for TBI, which may, potentially, induce negative side effects, such as social stigmatization or over-use with disengagement from other activities (Odom, et al., 2015).

Reliability. We reviewed the use of standardized vs. non-standardized and objective vs. subjective measurements across the set of studies. Since TBI evaluations often involve the use of several measurements, we counted the occurrence of each group of measurements across studies: 17 standardized measurements (12 objective and 5 subjective) and 37 non-standardized measurements (23 objective and 14 subjective). Figure 2 depicts the repartition of measurement groups according to the studies' purposes.

It is noteworthy that TE studies used more often standardized measurements (37.5%) than TU studies (11.1%). In contrast, both families of studies made similar use (roughly, 23%) of the least reliable measurements (*i.e.*, non-standardized subjective measurements). TE studies used therefore more reliable measurements than TU studies. Non-standardized measurements were dominant in TU studies (88.9%), with a majority of objective measurements (66.7%, Fig. 2). TU was often probed using dedicated, technology-related measurements to assess user accuracy (effectiveness), as well as yield (efficiency). For example, Zheng, Warren, et al. (2016) compared the performance between ASD and typically-developed participants on the level of prompting needed and the time spent to hit the target. It is unfortunate that the standardized methods for building usability measurements, such as Goal Attainment Scaling (GAS; Turner-Stokes, 2009) used by Valadão, et al. (2016), are not more widely used in TU studies. More surprisingly, none of the TU studies included the well-known, standardized

questionnaires for screening user-technology interactions or user experience, such as the SUS (Brooke, 1996) or QUEST 2.0, specially designed for children with disabilities (Demers, Weiss-Lambrou & Ska, 2000). As in TU studies, non-standardized measurements were dominant in TE-TU studies. Only one study (Fage, 2015) used standardized subjective tests for evaluating the effects of intervention, with the EQCA-VS (Morin & Maurice, 2001) for evaluating maladaptive behaviors (*i.e.*, TE outcome) and the USE questionnaire (Lund, 2001) for screening the TU (*i.e.*, technology usability and users' satisfaction).

Although a majority of studies addressed similar TBI outcomes, they often used different measurements for evaluating intervention effects. As a result, only two pairs of studies were found to share a standardized outcome measurement. Rice, et al. (2015) and Young & Posselt (2012) both used the Affect Recognition NEPSY subtest (Korkman, Kirk & Kemp, 2007) for evaluating emotion recognition. Rodriguez & Cummings (2016) and Whalen, et al. (2010) evaluated language abilities with the EVT (Williams, 1997) and PPVT (Dunn & Dunn, 2007). Shared endpoints across studies may support an accurate comparison of intervention effects across TBI. In turn, such comparison may move the field forward for identifying the best TBI for individuals with ASD, and even providing specific recommendations according to the ASD profile.

Durability and generalization of TBI effects: Near/far effects and transfer of acquired skills

Surprisingly, the durability of TBI effects has rarely been investigated, with only two TE studies assessing near/far effects (Grossman, Peskin & San Juan, 2013; Jeong, et al., 2015). A few weeks after the intervention (4 in Jeong, et al., 2015; 6-8 in Grossman, Peskin & San Juan, 2013), participants in both studies performed the same tasks as in the immediate post-intervention assessment. Performance results were similar in immediate and delayed post-tests, indicating that TBI effects, *i.e.* enhanced communication clarity (Grossman, Peskin & San Juan, 2013) and larger emotional vocabulary (Jeong, et al., 2015), were maintained after the intervention.

Eight TE studies included generalization measurements. Two TE studies used a standardized objective, but non-ecological measurement (*i.e.*, Happé's Strange Stories; Happé, 1994) to assess the transfer of target skills (Bauminger-Zviely, et al., 2013; Silver & Oakes, 2001). Three studies used standardized subjective measurements: parent-/teacher-reported measurements about real-life situations (*i.e.*, SSRS, Hopkins, et al., 2011; SRS, Rice, et al., 2015; SCQ, Young & Posselt, 2012). These data were used to examine the ecological transfer of TBI effects to social abilities in daily-life situations. Four studies used non-standardized measurements. This included two objective, hand-made tasks, where the participant had to apply newly-acquired skills in life-like situations (Grossman, Peskin & San Juan, 2013; Golan, et al., 2010). Generalization has also been assessed using non-standardized, subjective measurements to investigate the ecological transfer of skills in

real settings (Teachers' interviews, Lorenzo, et al., 2016; social interactions observations, Hopkins, et al., 2011 and Rice, et al., 2015).

In summary, near- and far-effects were rarely investigated since only two studies included a follow-up assessment for examining maintenance effects. Eight TE studies included an assessment of generalization, but only 6 of them used an ecological measurement. None of TU or TE-TU studies included either generalization or follow-up assessment. However, TU research may benefit from these aspects in the evaluation process of TU. First, maintenance effects may inform on the long-term usability experience of one product and then on its potential adoption by users. For instance, learning effects and expertise development may influence the users' needs, which in turn, will impact the product's usability. Second, generalization also deserves to be investigated for TU purposes for informing possible context-related variability that may impact the TU. The issue of "TU transfer" to real-life settings could be raised if we consider that a majority of TU studies are implemented in a laboratory, with a limited time of use.

Results consistency and its relationship with design and measurements

We reviewed methodological characteristics of studies addressing the TE and/or the TU of TBI with children and adolescents with ASD. This information is now linked with the evidence from studies, for examining the impacts of the methodology robustness on the reported TBI effects. In this section, we will review the evidence from included studies with respect to their *statistical* vs. *practical* significance (Ellis, 2010). The former is related to the significance of TBI effect on a given measure. The latter is related to the TBI effect size that can be assessed with Cohen's *d*.

For examining the *statistical* significance of TBI effect, included studies have been classified according to three levels: 1) highly-positive (significant TBI effects reported for all the outcomes within a study), 2) slightly-positive (mixed significance of TBI effects reported across the outcomes within a study), or 3) limited (moderate to non-significant TBI effects within a study) for each of the 31 studies reviewed (Table 5). Overall, TE studies reported inconsistent results concerning the TBI effect, *i.e.*, 7 with highly-positive, 8 with slightly-positive, and 8 with limited evidence. Fewer of the TBI effects reported in RCT studies were highly-positive (N= 3/14) than in controlled studies (N= 4/8, Table 5). Although there were fewer TU studies, all controlled trials, the TBI effects reported were mostly slightly-positive (N= 4/6). Hence, the highly-positive evidence for TBI was dependent on the study design, irrespective of its aim (TE *vs.* TU): the more robust the study design, the less consistent the results. Results' consistency was also related to the measurement reliability of TBI effects in both TE and TU studies (Fig. 3). First, standardized measurements yielded less consistent

evidence for a positive TBI effect (N= 8/16) than non-standardized ones (N= 18/23), irrespective of the aim of the study (TE *vs.* TU) (Fig. 3). Second, standardized measurements in TE studies were often associated with an RCT design (N= 8/14) but most of them showed moderately consistent evidence of TBI benefits (N= 11/14) (Table 5). In contrast, non-standardized measurements were frequently used in controlled trials (N= 7/8) and indicated highly-positive benefits of TBI (N= 4/7). TE studies with stricter methodological standards, an RCT design, and more reliable measurements produced less clear-cut evidence in favor of TBI than studies with a less-robust design and less-reliable measurements. Similarly, TU studies with less strict methodological standards also provided slightly- to highly-positive evidence of a TBI effect. The lack of standardized measurements associated with highly-positive results raises the issue of the reality of these TBI effects. Finally, among non-standardized measurements, subjective measurements were less frequently used in all studies (Fig. 2) and less associated with consistent positive evidence (N= 4/12) of a TBI effect (Fig. 3). This lack of consistency may be explained by the well-known biases of subjective rating (*i.e.*, self-assessment reliability or inter-rater reliability) (Annett, 2002).

Regarding *practical* significance of TBI effect, of the 31 studies, three studies did not report minimal data required for computing Cohen's *d* (Fage, 2015 [TE-TU]; Jeong, et al., 2015 [TE]; Valadão, et al., 2016 [TU]). Among the remaining studies, effect sizes ranged from -0.86 to 2.05 (details in Table 5). According to the Cohen's interpretation standards (Cohen, 1988), five TE studies resulted in small effect sizes ($0.2 < d < 0.5$), five in medium effect sizes ($0.5 < d < 0.7$), eight in large effect sizes ($d > 0.7$), and five TE studies yield none effect ($d < 0.2$). Two TU and two TE-TU studies yielded large effect sizes ($d > 0.7$), while three TU studies resulted in none effects ($d < 0.2$). Hence, the size of TBI effects did not appear to be linked to the study purpose (TE *vs.* TU). We did not observe relationships between the study design and the size of TBI effects: both controlled trials and RCTs yielded none to large effects. This result is not surprising, since study designs are more likely to affect the statistical significance than the practical significance: as seen earlier, the more robust the study design, the less consistent the results. For instance, the studies of Golan, et al. (2010) and Sitdhisanguan, et al. (2012) exhibited large effect sizes, despite of their differences in design (RCT *vs.* controlled trials, large *vs.* small sample sizes) and measurements (standardized *vs.* non-standardized).

However, we find a relationship between the measurement reliability and the size of TBI effects. With 0.7 as a threshold for large effect sizes (Table 5, bold values), studies with standardized measurements exhibited fewer large effect sizes (N= 3/9) than studies with non-standardized measurements (N= 11/19). Hence, the size of TBI effects appeared to be negatively related to the measurement reliability. This negative relationship suggested a well-known psychometric effect: non-standardized measurements may artificially inflate the

statistical significance as well as the practical significance, *i.e.*, effect size. In other words, the greater effect sizes reported in studies with hand-made measurements may be related to a psychometric bias due to the lack of measurement reliability. For example, both Golan, et al. (2010) and Jeong, et al. (2015) have measured the TBI outcome with a hand-made emotional vocabulary test. As the measurement reliability is not ensured, the change in the measure cannot be reliably associated with a real TBI effect. The lack of measurement reliability then compromises the generalization of the results to the emotional lexicon and even to emotional skills, which are the core targets in these studies.

Discussion

As previous reviews had already highlighted the study-design weaknesses in TBI literature for children and adolescents with ASD, this systematic review was restricted to 31 studies with the most robust designs.

The first stage was to examine the scope of TBI research and compare with previous findings. First, TBI studies were widely conducted with children (21), rather than adolescents (5). This agrees with Odom, et al. (2015), who observed the paucity of studies targeting adolescents with ASD. Further studies should address the late childhood and teen years for covering their support needs. Studies mostly involved computer- and robot-based interventions (19 studies). The large number of computer-based interventions was consistent with previous review (*e.g.*, Ploog, et al., 2013; Ramdoss, Lang, et al., 2011; Ramdoss, Mulloy, et al., 2011; Ramdoss, et al., 2012), while more robot-based studies were included than in previous reports (Grynszpan, et al., 2014). Robot-based interventions were therefore revealed as a new research trend in the field of TBI for ASD, given the particular interest in robotics among the ASD population and the robots' humanoid appearance (Begum, Serna & Yanco, 2016). This avenue of TBI research has received a growing attention, which might result in studies of greater quality than the studies surveyed in previous reviews (eight robot-based studies of our set were published between 2015 and 2016). However, robot-based interventions yielded less positive results than computer-based interventions. The reviewed TBIs mainly targeted emotional and/or social skills (15 studies) related to ASD. This fits with Grynszpan, et al. (2014), where 14 out of 21 studies included targeted socio-emotional skills. The review by Ramdoss, et al. (2012), targeting socio-emotional skills, also included a similar number of studies (12).

The next stage involved an examination of studies' methodology for assessing the TBI, with respect to the study purpose (*i.e.*, TE and/or TU). After the examination of study designs' characteristics, TE and TU studies were reviewed in depth for the reliability, consistency, generalization, and durability of TBI measurements, to obtain an accurate assessment of evidence-based practice standards. The result of this examination offers insight into study methods for clinical *vs.* ergonomic purposes. TE studies applied stricter methodological standards than TU studies, particularly in terms of study design, sample size, and inclusion/exclusion criteria. The examination of measurement reliability also supported the distinction between TE and TU studies. TE studies used more reliable measurements (*i.e.*, standardized ones), while TU studies made a large use of objective non-standardized measurements. However, all studies assessing TBI with individuals with ASD have to improve their design for taking account of external validity, durability and generalization of TBI effects.

TE studies provide evidence from promising to effective (levels 2 and 3), while TU study results range from emerging to promising (levels 1 and 2), according to the typology for classifying intervention studies by level of scientific evidence (Brownson, Fielding & Maylahn, 2009). This conclusion deserves consideration in future systematic reviews. As TE is usually the primary health interest of this type of review, it is critical to distinguish TE from TU studies due to their methodological differences. If the aim is to assess TE, the inclusion of TU studies in review data set may distort the results for TBI, as they are based on a less rigorous methodology (study design and measurements).

This conclusion is strengthened by our observations regarding the *statistical* and *practical* significance (*i.e.*, classification as highly-, slightly-positive and limited evidence and Cohen's *d*). Regarding *statistical* significance, studies with more robust designs elicited more inconsistent results (slightly-positive to limited evidence), while studies with less stringent designs yielded more frequently highly-positive evidence. For the *practical* significance, studies using less reliable measurements (*i.e.* non-standardized) more frequently elicited large effect sizes, while studies using standardized measurements elicited smaller effect sizes. The methodology robustness has therefore a real incidence on the results on TBI effects reported in studies amongst individuals with ASD. This observation may be harmful since most of studies on TBI suffer from methodological weaknesses and are incline to overestimate TBI effects. The large use of non-standardized measurements again distorts the evidence in assessing the TBI effects. Such measurements impede to reliably appreciate therapeutic benefits and can be confusing for the clinical interpretation such as the risks of a biased estimate of benefits-cost ratio of a TBI. As already recommended by Ramdoss, et al. (2012), hand-made measurements have to be standardized if there are considered useful for capturing the outcomes of an intervention. The procedure of standardization might also elucidate the question of correspondence between score and real-life outcomes and allow reliably assessing the TBI effects.

Three TE studies had a methodologically sound study design (*i.e.*, RCT) and measurements (*i.e.*, standardized) (Hopkins, et al., 2011; Rice, et al., 2015 and Young & Posselt, 2012). They shared a similar clinical purpose, *i.e.*, to improve the socio-emotional abilities of children with ASD, like many TBI studies for ASD (*e.g.*, Grynszpan, et al., 2014; Ramdoss, et al., 2012). Interestingly, two of these studies conducted a TBI using the same computer program (*FaceSay*®): Rice, et al. (2015) extended the results of Young & Posselt (2012). Hopkins, et al. (2011) also extended previous results, using *The Transporters*® DVD as TBI for children with ASD. These interventions were conducted at school or at home (similar to real-life settings) for a period ranging from 2 to 10 weeks. Internal validity was respected with a good reliability, since the studies used standardized measurements to assess direct outcomes; whereas, like all the studies in our dataset, external

validity was not investigated. Durability was not studied, but all three RCTs included an ecological transfer assessment, involving standardized subjective tests on social skills in real settings (*e.g.*, SSRS, Gresham & Elliot, 1990; SCQ, Rutter, Bailey & Lord, 2003). These three studies were classified as slightly-positive because they did not report significant positive evidence for all outcomes. However, they elicited effect sizes reflecting small to large effect, which account for the promising aspects of TBI with children with ASD.

TE and TU Studies – friends or foes?

Only 3 studies addressed both TE and TU to validate their TBI (Fage, 2015; Fage, et al., 2016; Sitdhisanguan, et al., 2012). Unlike TU studies, very little research investigating both issues was found in our initial search. This may indicate that, today, TE and TU are not considered two complementary dimensions in the TBI domain for ASD. However, TE and TU are complementary facets, which deserve to be investigated simultaneously in TBI studies. It is methodologically relevant that a TBI study should cover both TU and TE aspects to document TBI-related uses and usages, as well as health benefits, and even the relationships between these factors. The three TE-TU studies were less rigorous than most TE studies but used standardized or objective measurements. They also addressed internal, but not external, validity, as well as durability and generalization. These studies represent a promising research approach for TBI investigation in children and adolescents with ASD, by combining ergonomic and clinical results for an in-depth investigation of TBI effects. They attempted to provide a trade-off between the advantages of both health and ergonomics research. However, further studies should make effort to apply evidence-based practice standards (sample size, study design, study measurements) to reinforce this promising, emergent approach.

Considering that TE and TU may be complementary facets, the distinction between TE, TU and TE-TU studies offers perspectives for further research in the field of TBI with ASD.

On the one hand, all TE study should consider TU as a pre-requisite for the therapeutic benefits of any TBI. A TBI may elicit substantial therapeutic benefits only if the product is usable for the targeted users. Hence, TU examination deserves consideration when inspecting the TE of one TBI. The examination of drop-outs and the reasons why across studies may be informative of such consideration. For instance, Gordon, et al. (2014) excluded eleven participants due to unusable data, leading to put into question the TU of this TBI. In contrast, some TE studies have taken into account TU recommendations for designing TBI for individuals with ASD and reported design guidelines in their article (*e.g.*, Bartoli, et al., 2014). These guidelines may inform on main TU issues experienced with individuals with ASD but cannot replace a TU assessment. Another way to guarantee the TU of a TBI may be found in using participatory design frameworks, which include future users from the

beginning of the design process. These design methods help to maximize the TU of one product and implies several TU assessments during the design process.

On the other hand, TU studies added a contribution to the field of TBI for individuals with ASD by screening the users' needs and issues relating to the technology. However, TU studies may be improved with a greater consideration of TE when assessing a TBI. As it is, TU studies do not permit to recommend TBI for individuals with ASD. They provide evidence for the usability of a TBI but remain of little clinical usefulness because TE is not reliably addressed. A practitioner looking for a useful TBI for a patient with ASD will rely on clinical evidence in order to preconize a TBI that effectively address the patients' needs. More than being usable, TBI have to demonstrate evidence for a substantive gain in daily lives of people with ASD. Then, TU should at least specify that TE benefits are not fully addressed and that further studies are needed for reliably accounting for the TBI clinical usefulness. This is of greater importance when we consider common public expectations on TBI research for supporting individuals with ASD.

Limitations

This systematic review has several limitations. First, only one coder conducted the study search, which is, by definition, a limitation on this review. To test of inter-rater reliability during data selection, two researchers independently applied the Jadad/SIGN criteria in a review of 8 randomly-selected articles. Both researchers met to discuss their differences and reach a consensus on the application of the Jadad/SIGN criteria. One researcher conducted data analysis for the remaining articles. Any doubts were discussed before excluding studies. The same procedure was applied during study analysis for the criteria relative on reliability, consistency, generalization, and durability effect. During the entire systematic review procedure, both researchers met several times to check the observance of criteria lists, by comparing and reconciling differences.

The set of articles (N= 31) overlapped very little with previous reviews (*i.e.*, four with Ploog, et al., 2013; three with Knight, McKissick & Saunders, 2013; two with Odom, et al., 2015; and four with Grynszpan, et al., 2014). Furthermore, a large majority of studies (22) investigated TE issues, while only 6 focused on TU. These discrepancies may be ascribed to two major reasons: the search process or the inclusion/exclusion criteria.

First, even if we based our search process on PICO criteria, some studies may have fallen between the cracks. For instance, the small number of TU and TE-TU studies may be attributed to a problem with the search query, rather than their absence from the literature. The PICO method is well adapted to research in the health intervention field but may be less sensitive for usability studies' screening. Other alternatives would be possible, such as SPIDER, presented as a better tool than PICO (Cooke, Smith & Booth, 2012). However, a comparative

study of PICO and SPIDER showed that PICO was more sensitive and SPIDER more specific. The authors finally recommended the use of PICO to compensate for the lesser sensitivity of SPIDER (Methley, et al., 2014).

Second, our inclusion and exclusion criteria may be too severe, leading to the drastic pruning before applying SIGN ratings (204 excluded). The most frequent reason for exclusion at this stage was the use of single-case designs. This fits with the large proportion of such studies included in previous reviews (e.g., Knight, McKissick & Saunders, 2013; Odom, et al., 2015). However, Knight, McKissick & Saunders (2013) raised concerns about the validity of such studies, since only 4/17 single-case design studies of their set were considered of "acceptable" quality. Grynszpan, et al. (2014) also excluded this kind of design in their meta-analysis. The next selection stage reduced again the number of included articles (48/79 excluded). The application of SIGN ratings mainly excluded non-comparative studies and might explain the discrepancy with the studies' set of Ploog, et al. (2013) for instance. The presence of a control group prevents the results from the effects of growth and cognitive development that are likely to interfere with intervention effects. This is of greater importance when we consider that TBI mainly targeted children with ASD, which are characterized by a large heterogeneity. Another point is that our intention was to review the literature with concerns to the standards of evidence-based practices. TBI have to provide the highest evidence of their efficacy for being prescribed to children with ASD. Yet, the gold standards for validating a therapeutic technique is to conduct RCTs. We first reviewed the literature with the willing to only include RCTs, and finally enlarged our criteria to controlled studies. These latter have the potential to assess intervention effects with a great level of evidence when they are well conducted. Regarding evidence-based practices, controlled studies are related to the minimal level of evidence, while RCT has a greater value. The drastic pruning during the selection process may inform that we still are far from recognizing TBI as evidence-based practices with individuals with ASD. Further studies should strengthen their design and consider the use of standardized measurements for reliably valuating the TBI effects.

To conclude, the present systematic review identified some methodological flaws in the research field of TBI for children and adolescents with ASD. Although a number of well-conducted studies reported promising results, we must be careful not "to throw the baby out with the bath water" by trying to learn from the best and to end up with the worst. As an emerging interdisciplinary TBI research approach, studies addressing both TE and TU might provide fruitful approach by combining expertise in human-computer interaction and health research for yielding methodological empowerment.

References

References marked with an asterisk indicate studies included in the systematic review.

Agarwal, A., & Meyer, A. (2009). Beyond Usability: Evaluating Emotional Response as an Integral Part of the User Experience. Paper presented at *the 27th Annual ACM Conference on Human Factors in Computing Systems (CHI EA 2009), Boston, MA, USA* (pp. 2919–2930). http://doi.org/10.1145/1520340.1520420

American Psychiatric Association [APA]. (2013). *Diagnostic and Statistical Manual of Mental Disorders: DSM-V* (5th ed.). Washington, DC, USA: APA.

Annett, J. (2002). Subjective Rating Scales: Science or Art? *Ergonomics*, *45*(14), 966-987. http://dx.doi.org/10.1080/00140130210166951

Ardoin, S. P. (2006). The Response in Response to Intervention: Evaluating the Utility of Assessing Maintenance of Intervention Effects. *Psychology in the Schools*, *43*(6), 713-725. http://doi.org/10.1002/pits.20181

Baharuddin, R., Singh, D., & Razali, R. (2013). Usability Dimensions for Mobile Applications: A Review. *Research Journal of Applied Sciences, Engineering and Technology*, *5*(6), 2225–2231. Retrieved from http://maxwellsci.com/jp/abstract.php?jid=RJASET&no=266&abs=55

*Bartoli, L., Garzotto, F., Gelsomini, M., Oliveto, L., & Valoriani, M. (2014). Designing and Evaluating Touchless Playful Interaction for ASD Children. Paper presented at *the 8th Conference on Interaction Design and Children (IDC 2014), Aarhus, Denmark* (pp. 17-26). http://doi.org/10.1145/2593968.2593976

*Bauminger-Zviely, N., Eden, S., Zancanaro, M., Weiss, P. L. T., & Gal, E. (2013). Increasing Social Engagement in Children with High-Functioning Autism Spectrum Disorder Using Collaborative Technologies in the School Environment. *Autism*, *17*(3), 317–339. http://doi.org/10.1177/1362361312472989

Beery, K. E., & Beery, N. A. (2004). *The Beery-Buktenica Developmental Test of Visual-Motor Integration (5th ed.)* Minneapolis, MN, USA: NCS Pearson.

Begum, M., Serna, R. W., & Yanco, H. A. (2016). Are Robots Ready to Deliver Autism Interventions? A Comprehensive Review. *International Journal of Social Robotics*, *8*(2), 157–181. http://doi.org/10.1007/s12369-016-0346-y

*Bekele, E. T., Crittendon, J. A., Swanson, A. R., Sarkar, N., & Warren, Z. E. (2014). Pilot Clinical Application of an Adaptive Robotic System for Young Children with Autism. *Autism*, *18*(5), 598–608. http://doi.org/10.1177/1362361313479454

*Bekele, E. T., Zheng, Z., Swanson, A. R., Crittendon, J. A., Warren, Z. E., & Sarkar, N. (2013). Understanding How Adolescents with Autism Respond to Facial Expressions in Virtual Reality Environments. *IEEE*

Transactions on Visualization and Computer Graphics, *19*(4), 711–720. http://doi.org/10.1109/TVCG.2013.42

Bevan, N., Carter, J., & Harker, S. (2015). ISO 9241-11 Revised: What Have We Learnt About Usability Since 1998? In M. Kurosu (Ed.), *Human-Computer Interaction - Design and Evaluation: 17th International Conference, HCI International 2015* (LNCS Vol. 9169, pp. 143–151). http://doi.org/10.1007/978-3-319-20901-2_13

Brooke, J. (1996). SUS - A Quick and Dirty Usability Scale. *Usability Evaluation in Industry*, *189*(194), 4–7. Retrieved from http://hell.meiert.org/core/pdf/sus.pdf

Brownson, R. C., Fielding, J. E., & Maylahn, C. M. (2009). Evidence-Based Public Health: A Fundamental Concept for Public Health Practice. *Annual Review of Public Health*, *30*(1), 175–201. http://doi.org/10.1146/annurev.publhealth.031308.100134

Chan, R. C. K., Shum, D., Toulopoulou, T., & Chen, E. Y. H. (2008). Assessment of Executive Functions: Review of Instruments and Identification of Critical Issues. *Archives of Clinical Neuropsychology*, *23*(2), 201–216. http://doi.org/10.1016/j.acn.2007.08.010

Cohen, J. (1988). *Statistical Analysis for the Behavioral Sciences, 2nd Edition.* Hillsdale, NJ: Lawrence Erlbaum.

Constantino, J. N., & Gruber, C. P. (2005). *Social responsiveness scale (SRS)* (Western Ps). Los Angeles, CA,USA.

Cooke, A., Smith, D., & Booth, A. (2012). Beyond PICO: The SPIDER Tool for Qualitative Evidence Synthesis. *Qualitative Health Research*, *22*(10), 1435–1443. http://doi.org/10.1177/1049732312452938

*Costescu, C. A., Vanderborght, B., & David, D. O. (2015). Reversal Learning Task in Children with Autism Spectrum Disorder: A Robot-Based Approach. *Journal of Autism and Developmental Disorders*, *45*(11), 3715–3725. http://doi.org/10.1007/s10803-014-2319-z

Demers, L., Weiss-Lambrou, R., & Ska, B. (2000). Item Analysis of the Quebec User Evaluation of Satisfaction with Assistive Technology (QUEST). *Assistive Technology*, *12*(2), 96–105. http://doi.org/10.1080/10400435.2000.10132015

Drost, E. A. (2011). Validity and Reliability in Social Science Research. *Education Research and Perspectives*, *38*(1), 105–123. Retrieved from http://erpjournal.net/wp-content/uploads/2012/07/ERPV38-1.-Drost-E.-2011.-Validity-and-Reliability-in-Social-Science-Research.pdf

Dunn, D. M., & Dunn, L. M. (2007). *Peabody Picture Vocabulary Test: Manual*. San Antonio, TX, USA: Pearson, Inc.

Ekman, P., & Friesen, W. (1976). *Pictures of Facial Affect*. Palo Alto, CA, USA: Consulting Psychologists Press.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.

*Fage, C. (2015). An Emotion Regulation App for School Inclusion of Children with ASD: Design Principles and Preliminary Results for Its Evaluation. *ACM SIGACCESS Accessibility and Computing Newsletter*, (112), 8–15. http://doi.org/10.1145/2809915.2809917

*Fage, C., Pommereau, L., Consel, C., Balland, E., & Sauzéon, H. (2016). Tablet-Based Activity Schedule in Mainstream Environment for Children with Autism and Children with ID. *ACM Transactions on Accessible Computing (TACCESS)*, *8*(3), 1–26. http://doi.org/10.1145/2854156

*Falkmer, T., Horlin, C., Dahlman, J., Dukic, T., Barnett, T., & Anund, A. (2014). Usability of the SAFEWAY2SCHOOL System in Children with Cognitive Disabilities. *European Transport Research Review*, *6*(2), 127–137. http://doi.org/10.1007/s12544-013-0117-x

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2. http://doi.org/10.1037/a0024338

*Golan, O., Ashwin, E., Granader, Y., McClintock, S., Day, K., Leggett, V., & Baron-Cohen, S. (2010). Enhancing Emotion Recognition in Children with Autism Spectrum Conditions: An Intervention Using Animated Vehicles with Real Emotional Faces. *Journal of Autism and Developmental Disorders*, *40*(3), 269–279. http://doi.org/10.1007/s10803-009-0862-9

*Gordon, I., Pierce, M. D., Bartlett, M. S., & Tanaka, J. W. (2014). Training Facial Expression Production in Children on the Autism Spectrum. *Journal of Autism and Developmental Disorders*, *44*(10), 2486–2498. http://doi.org/10.1007/s10803-014-2118-6

Gresham, F. M., & Elliot, S. N. (1990). *Social Skills Rating System Manual*. Circle Pines, MN, USA: American Guidance Service.

Grondin, S. C., & Schieman, C. (2011). Evidence-Based Medicine: Levels of Evidence and Evaluation Systems. In M. K. Ferguson (Ed.), *Difficult Decisions in Thoracic Surgery* (pp. 13–22). http://doi.org/10.1007/978-1-84996-492-0_2

*Grossman, M., Peskin, J., & San Juan, V. (2013). Thinking About a Reader's Mind: Fostering Communicative Clarity in the Compositions of Youth with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, *43*(10), 2376–2392. http://doi.org/10.1007/s10803-013-1786-y

*Grynszpan, O., Martin, J.-C., & Nadel, J. (2008). Multimedia Interfaces for Users with High Functioning Autism: An Empirical Investigation. *International Journal of Human-Computer Studies*, *66*(8), 628–639. http://doi.org/10.1016/j.ijhcs.2008.04.001

Grynszpan, O., Weiss, P. L. T., Perez-Diaz, F., & Gal, E. (2014). Innovative Technology-based Interventions for Autism Spectrum Disorders: A Meta-analysis. *Autism*, *18*(4), 346–361. http://doi.org/10.1177/1362361313476767

Happé, F. G. E. (1994). An Advanced Test of Theory of Mind: Understanding of Story Characters' Thoughts and Feelings by Able Autistic, Mentally Handicapped, and Normal Children and Adults. *Journal of Autism and Developmental Disorders*, *24*(2), 129–154. http://doi.org/10.1007/BF02172093

Hayes, G. R., Hirano, S., Marcu, G., Monibi, M., Nguyen, D. H., & Yeganyan, M. (2010). Interactive Visual Supports for Children with Autism. *Personal and Ubiquitous Computing*, *14*(7), 663–680. http://doi.org/10.1007/s00779-010-0294-8

Hersh, M. (2014). Evaluation Framework for ICT-based Learning Technologies for Disabled People. *Computers & Education*, *78*, 30–47. http://doi.org/10.1016/j.compedu.2014.05.001

Holzinger, A. (2005). Usability engineering methods for software developers. *Communications of the ACM, 48*(1), 71-74. http://doi.org/10.1145/1039539.1039541 .

*Hopkins, I. M., Gower, M. W., Perez, T. A., Smith, D. S., Amthor, F. R., Wimsatt, F. C., & Biasini, F. J. (2011). Avatar Assistant: Improving Social Skills in Students with an ASD Through a Computer-Based Intervention. *Journal of Autism and Developmental Disorders*, *41*(11), 1543–1555. http://doi.org/10.1007/s10803-011-1179-z

Hourcade, J. P., Williams, S. R., Miller, E. A., Huebner, K. E., & Liang, L. J. (2013). Evaluation of Tablet Apps to Encourage Social Interaction in Children with Autism Spectrum Disorders. In W. E. Mackay, S. Brewster, & S. Bødker (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'13), Paris, France* (pp. 3197–3206). http://doi.org/10.1145/2470654.2466438

Hutchins, T. L., Prelock, P. A., & Bonazinga, L. (2012). Psychometric Evaluation of the Theory of Mind Inventory (ToMI): A Study of Typically Developing Children and Children with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *42*(3), 327–341. http://doi.org/10.1007/s10803-011-1244-7

Inostroza, R., Rusu, C., Roncagliolo, S., & Rusu, V. (2013). Usability Heuristics for Touchscreen-based Mobile Devices: Update. Paper presented at *the 2013 Chilean Conference on Human - Computer Interaction (ChileCHI2013), Temuco, Chile*, (pp. 24-29). http://doi.org/10.1145/2535597.2535602

International Organisation for Standardisation (ISO). (2014). *Guide for addressing accessibility in standards*. ISO/IEC Guide 71:2014. Geneva, Switzerland : ISO.

Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J. M., Gavaghan, D. J., … Relief, P. (1996). Assessing the Quality of Reports of Randomized Clinical Trials : Is Blinding Necessary ? *Controlled Clinical Trials*, *17*, 1–12. http://doi.org/10.1016/0197-2456(95)00134-4

*Jeong, M., Kim, Y., Yim, D., Yeon, S., Song, S., & Kim, J. (2015). Lexical Representation of Emotions for High Functioning Autism (HFA) via Emotional Story Intervention Using Smart Media. Paper presented at *the 33rd Annual ACM Conference on Human Factors in Computing Systems – Extended Abstracts (CHI EA 2015), Seoul, Republic of Korea* (pp. 1983–1988). http://doi.org/10.1145/2702613.2732750

Kenworthy, L., Yerys, B. E., Anthony, L. G., & Wallace, G. L. (2008). Understanding Executive Control in Autism Spectrum Disorders in the Lab and in the Real World. *Neuropsychology Review*, *18*(4), 320–338. http://doi.org/10.1007/s11065-008-9077-7

Knight, V. F., McKissick, B. R., & Saunders, A. (2013). A Review of Technology-Based Interventions to Teach Academic Skills to Students with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *43*(11), 2628–2648. http://doi.org/10.1007/s10803-013-1814-y

Korkman, M., Kirk, U., & Kemp, S. (2007). *NEPSY - Second Edition (NEPSY - II): A Developmental Neuropsychological Assessment*. San Antonio, TX, USA: Psychological Corporation.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H. J., Leventhal, B. L., DiLavore, P. C., … Rutter, M. (2000). The Autism Diagnostic Schedule – Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism. *Journal of Autism and Developmental Disorders*, *30*(3), 205–223. http://doi.org/10.1023/A:1005592401947

*Lorenzo, G., Lledó, A., Pomares, J., & Roig, R. (2016). Design and Application of an Immersive Virtual Reality System to Enhance Emotional Skills for Children with Autism Spectrum Disorders. *Computers & Education*, *98*, 192–205. http://doi.org/10.1016/j.compedu.2016.03.018

Lund, A. M. (2001). Measuring Usability with the USE questionnaire. *Usability Interface*, *8*(2), 3–6. http://doi.org/10.1177/1078087402250360

McCartney, K., & Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. *Child Development*, *71*(1), 173-180. https://doi.org/10.1111/1467-8624.00131

Methley, A. M., Campbell, S., Chew-Graham, C., McNally, R., & Cheraghi-Sohi, S. (2014). PICO, PICOS and SPIDER: A Comparison Study of Specificity and Sensitivity in Three Search Tools for Qualitative Systematic Reviews. *BMC Health Services Research*, *14*(579), 1–10. http://doi.org/10.1186/s12913-014-0579-0

Morin, D., & Maurice, P. (2001). Élaboration de la Version Scolaire de l'Echelle Quebecoise de Comportements Adaptatifs (ECQA-VS). *Revue Francophone de La Déficience Intellectuelle*, *12*(1), 7–20. Retrieved from http://www.rfdi.org/elaboration-de-la-version-scolaire-de-lechelle-quebecoise-de-comportements-adaptatifs-eqca-vs/

Odom, S. L., Thompson, J. L., Hedges, S., Boyd, B. A., Dykstra, J. R., Duda, M. A., … Bord, A. (2015). Technology-Aided Interventions and Instruction for Adolescents with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *45*(12), 3805–3819. http://doi.org/10.1007/s10803-014-2320-6

*Ploog, B. O., Banerjee, S., & Brooks, P. J. (2009). Attention to Prosody (Intonation) and Content in Children with Autism and in Typical Children Using Spoken Sentences in a Computer Game. *Research in Autism Spectrum Disorders*, *3*(3), 743–758. http://doi.org/10.1016/j.rasd.2009.02.004

Ploog, B., Scharf, A., Nelson, D., & Brooks, P. (2013). Use of Computer-Assisted Technologies (CAT) to Enhance Social, Communicative, and Language Development in Children with Autism Spectrum Disorders. *Journal of Autism & Developmental Disorders*, *43*(2), 301–322. http://doi.org/10.1007/s10803-012-1571-3

*Pop, C. A., Pintea, S., Vanderborght, B., & David, D. O. (2014). Enhancing Play Skills, Engagement and Social Skills in a Play Task in ASD Children by Using Robot-based Interventions. A Pilot Study. *Interaction Studies*, *15*(2), 292–320. http://doi.org/10.1075/is.15.2.14pop

*Pop, C. A., Simut, R. E., Pintea, S., Saldien, J., Rusu, A. S., Vanderfaeillie, J., … Vanderborght, B. (2013). Social Robots *vs.* Computer Display: Does the Way Social Stories are Delivered Make a difference for Their Effectiveness on ASD Children? *Journal of Educational Computing Research*, *49*(3), 381–401. http://doi.org/10.2190/EC.49.3.f

Putnam, C., & Chong, L. (2008). Software and Technologies Designed for People with Autism: What Do Users Want? Paper presented at *the 10th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2008), Halifax, Canada*, (pp. 3–10). http://doi.org/10.1145/1414471.1414475

Ramdoss, S., Lang, R., Mulloy, A., Franco, J., O'Reilly, M., Didden, R., & Lancioni, G. (2011). Use of Computer-Based Interventions to Teach Communication Skills to Children with Autism Spectrum Disorders: A Systematic Review. *Journal of Behavioral Education*, *20*(1), 55–76. http://doi.org/10.1007/s10864-010-9112-7

Ramdoss, S., Machalicek, W., Rispoli, M., Mulloy, A., Lang, R., & O'Reilly, M. (2012). Computer-based Interventions to Improve Social and Emotional Skills in Individuals with Autism Spectrum Disorders: A Systematic Review. *Developmental Neurorehabilitation*, *15*(2), 119–135. http://doi.org/10.3109/17518423.2011.651655

Ramdoss, S., Mulloy, A., Lang, R. B., O'Reilly, M. F., Sigafoos, J., Lancioni, G. E., … El Zein, F. (2011). Use of Computer-based Interventions to Improve Literacy Skills in Students with Autism Spectrum Disorders: A Systematic Review. *Research in Autism Spectrum Disorders*, *5*(4), 1306–1318. http://doi.org/10.1016/j.rasd.2011.03.004

Reed, P., & Osborne, L. A. (2014). Mainstream education for children with autism spectrum disorders. In Tarbox, J., Dixon, D., Sturmey, P. & Matson, J. (Eds). *Handbook of Early Intervention for Autism Spectrum Disorders.* Autism and Child Psychopathology Series. (pp. 447-485). Springer, New York, NY.

*Rice, L., Wall, C., Fogel, A., & Shic, F. (2015). Computer-Assisted Face Processing Instruction Improves Emotion Recognition, Mentalizing, and Social Skills in Students with ASD. *Journal of Autism and Developmental Disorders*, *45*(7), 2176–2186. http://doi.org/10.1007/s10803-015-2380-2

*Rodríguez, C. D., & Cumming, T. M. (2016). Employing Mobile Technology to Improve Language Skills of Young Students with Language-based Disabilities. *Assistive Technology*, Latest Articles, 1–9. http://doi.org/10.1080/10400435.2016.1171810

Rutter, M., Bailey, A., & Lord, C. (2003). *The Social Communication Questionnaire (SCQ)*. Los Angeles, CA, USA: Western Psychological Services.

*Salvador, M. J., Silver, S., & Mahoor, M. H. (2015). An Emotion Recognition Comparative Study of Autistic and Typically-developing Children Using the Zeno Robot. Paper presented at *the 2015 IEEE International Conference on Robotics and Automation (ICRA 2015), Seattle, WA, USA* (pp. 6128–6133). http://doi.org/10.1109/ICRA.2015.7140059

Scottish Intercollegiate Guidelines Network [SIGN]. (2008). *SIGN50: A Guideline Developer's Handbook.* Edinburgh: SIGN.

Sharafi, Z., Soh, Z., & Guéhéneuc, Y. G. (2015). A Systematic Literature Review on the Usage of Eye-tracking in Software Engineering. *Information and Software Technology*, *67*, 79–107. http://doi.org/10.1016/j.infsof.2015.06.008

*Silver, M., & Oakes, P. (2001). Evaluation of a New Computer Intervention to Teach People with Autism or Asperger Syndrome to Recognize and Predict Emotions in Others. *Autism*, *5*(3), 299–316. http://doi.org/10.1177/1362361301005003007

Simms, L. J. (2008). Classical and Modern Methods of Psychological Scale Construction. *Social and Personality Psychology Compass*, *2*(1), 414–433. http://doi.org/10.1111/j.1751-9004.2007.00044.x

*Sitdhisanguan, K., Chotikakamthorn, N., Dechaboon, A., & Out, P. (2012). Using Tangible User Interfaces in Computer-based Training Systems for Low-functioning Autistic Children. *Personal and Ubiquitous Computing*, *16*(2), 143–155. http://doi.org/10.1007/s00779-011-0382-4

*Srinivasan, S. M., Eigsti, I.-M., Gifford, T., & Bhat, A. N. (2016a). The Effects of Embodied Rhythm and Robotic Interventions on the Spontaneous and Responsive Verbal Communication Skills of Children with Autism Spectrum Disorder (ASD): A Further Outcome of a Pilot Randomized Controlled Trial. *Research in Autism Spectrum Disorders*, *27*, 73–87. http://doi.org/10.1016/j.rasd.2016.04.001

*Srinivasan, S. M., Eigsti, I.-M., Neelly, L. B., & Bhat, A. N. (2016b). The Effects of Embodied Rhythm and Robotic Interventions on the Spontaneous and Responsive Social Attention Patterns of Children with Autism Spectrum Disorder (ASD): A Pilot Randomized Controlled Trial. *Research in Autism Spectrum Disorders*, *27*, 54–72. http://doi.org/10.1016/j.rasd.2016.01.004

*Srinivasan, S. M., Park, I. K., Neelly, L. B., & Bhat, A. N. (2015). A Comparison of the Effects of Rhythm and Robotic Interventions on Repetitive Behaviors and Affective States of Children with Autism Spectrum Disorder (ASD). *Research in Autism Spectrum Disorders*, *18*, 51–63. http://doi.org/10.1016/j.rasd.2015.07.004

Taylor, J. L., Henninger, N. A., & Mailick, M. R. (2015). Longitudinal patterns of employment and postsecondary education for adults with autism and average-range IQ. *Autism*, *19*(7), 785-793. http://doi.org/10.1177/1362361315585643

Taylor, L. J., Eapen, V., Maybery, M. T., Midford, S., Paynter, J., Quarmby, L., … Whitehouse, A. J. O. (2016). Diagnostic Evaluation for Autism Spectrum Disorder: a Survey of Health Professionals in Australia. *BMJ Open*, *6*(9), e012517. http://doi.org/10.1136/bmjopen-2016-012517

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Practitioner Review: Do Performance-based Measures and Ratings of Executive Function Assess the Same Construct? *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *54*(2), 131–143. http://doi.org/10.1111/jcpp.12001

Turner-Stokes, L. (2009). Goal Attainment Scaling (GAS) in Rehabilitation: A Practical Guide. *Clinical Rehabilitation*, *23*(4), 362–70. http://doi.org/10.1177/0269215508101742

*Valadão, C. T., Goulart, C., Rivera, H., Caldeira, E., Bastos Filho, T. F., Frizera-Neto, A., & Carelli, R. (2016). Analysis of the Use of a Robot to Improve Social Skills in Children with Autism Spectrum Disorder. *Research on Biomedical Engineering*, *32*(2), 161-175. http://doi.org/10.1590/2446-4740.01316

Volkmar, F., Siegel, M., Woodbury-Smith, M., King, B., McCracken, J., & State, M. (2014). Practice Parameter for the Assessment and Treatment of Children and Adolescents with Autism Spectrum Disorder. *Journal*

of the American Academy of Child and Adolescent Psychiatry, *53*(2), 237–257. http://doi.org/10.1016/j.jaac.2013.10.013

Wechsler, D. (2003). Wechsler Intelligence Scale for Children-Fourth Version (WISC-IV) San Antonio, TX, USA: Psychological Corporation.

Wechsler, D. (2014). Wechsler Abbreviated Scale of Intelligence -Second Edition (WASI-II). San Antonio, TX, USA: Psychological Corporation.

*Whalen, C., Moss, D., Ilan, A. B., Vaupel, M., Fielding, P., Macdonald, K., … Symon, J. (2010). Efficacy of TeachTown: Basics Computer-assisted Intervention for the Intensive Comprehensive Autism Program in Los Angeles Unified School District. *Autism*, *14*(3), 179–197. http://doi.org/10.1177/1362361310363282

Williams, K. T. (1997). *The Expressive Vocabulary Test (EVT)*. Circle Pines, MN, USA: American Guidance Service.

*Young, R. L., & Posselt, M. (2012). Using The Transporters DVD as a Learning Tool for Children with Autism Spectrum Disorders (ASD). *Journal of Autism and Developmental Disorders*, *42*(6), 984–991. http://doi.org/10.1007/s10803-011-1328-4

*Zheng, Z., Warren, Z. E., Weitlauf, A. S., Fu, Q., Zhao, H., Swanson, A. R., & Sarkar, N. (2016a). Brief Report: Evaluation of an Intelligent Learning Environment for Young Children with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *46*(11), 3615–3621. http://doi.org/10.1007/s10803-016-2896-0

*Zheng, Z., Young, E. M., Swanson, A. R., Weitlauf, A. S., Warren, Z. E., & Sarkar, N. (2016b). Robot-Mediated Imitation Skill Training for Children With Autism. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *24*(6), 682–691. http://doi.org/10.1109/TNSRE.2015.2475724

Figure Captions

Figure 1. Flow diagram of studies' selection.

Figure 2. Measurements' repartition according to their type. Y-axis corresponds to the purpose of studies, and x-axis to the percentage values. Percentages are given according to each type of measures [*i.e.* striped grey: standardized objective (STD Obj.), striped black: standardized subjective (STD Subj.), plain grey: non-standardized objective (N-STD Obj.) and plain black: non-standardized subjective (N-STD Subj.)].

Figure 3. Percentage of full positive evidence reported according to results obtained for each single measure in studies. Percentages have been computed with respect to the type of measures. Y-axis corresponds to types of measures and results for all studies and TE studies. X-axis corresponds to the percentage value.

Figure 1 top

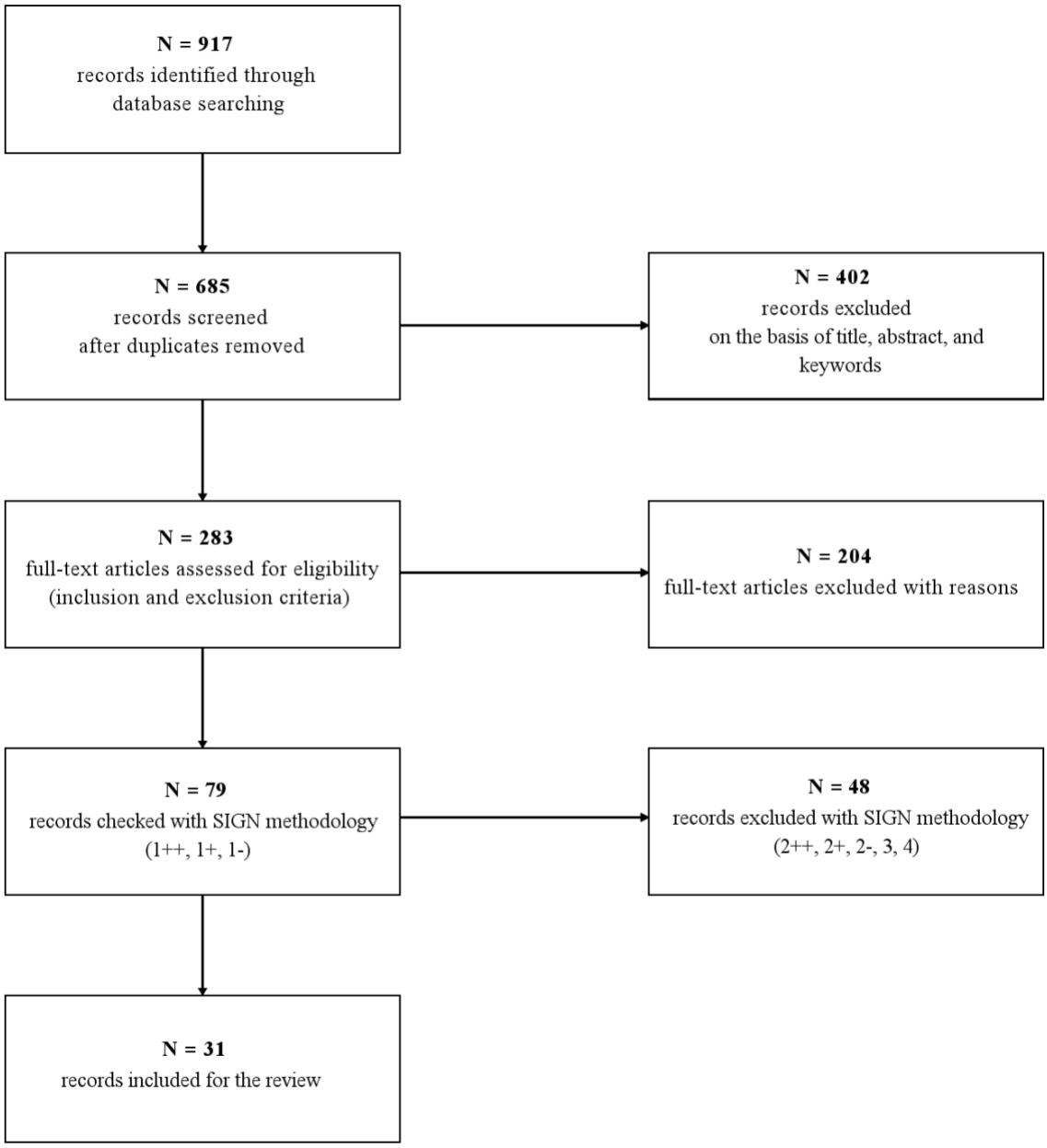
N = 917
records identified through
database searching
N = 685
records screened
after duplicates removed
N = 402
records excluded
on the basis of title, abstract, and
keywords
N = 283
full-text articles assessed for eligibility
(inclusion and exclusion criteria)
N = 204
full-text articles excluded with reasons
N = 79
records checked with SIGN methodology
(1++, 1+, 1-)
N = 48
records excluded with SIGN methodology
(2++, 2+, 2-, 3, 4)
N = 31
records included for the review

Figure 2 top

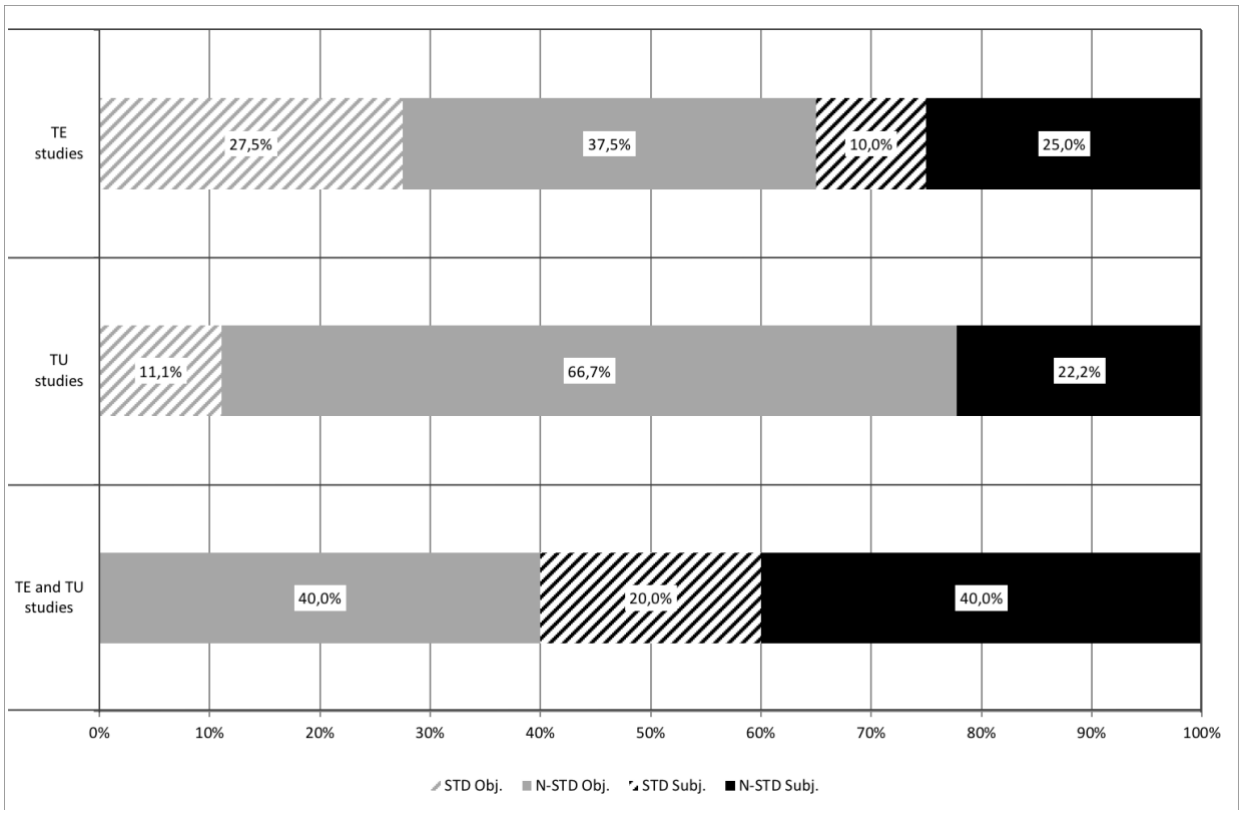
TE studies
27,5%
37,5%
10,0%
25,0%
TU studies
11,1%
66,7%
22,2%
TE and TU studies
40,0%
20,0%
40,0%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
STD Obj.
N-STD Obj.
STD Subj.
N-STD Subj.

Figure 3 top

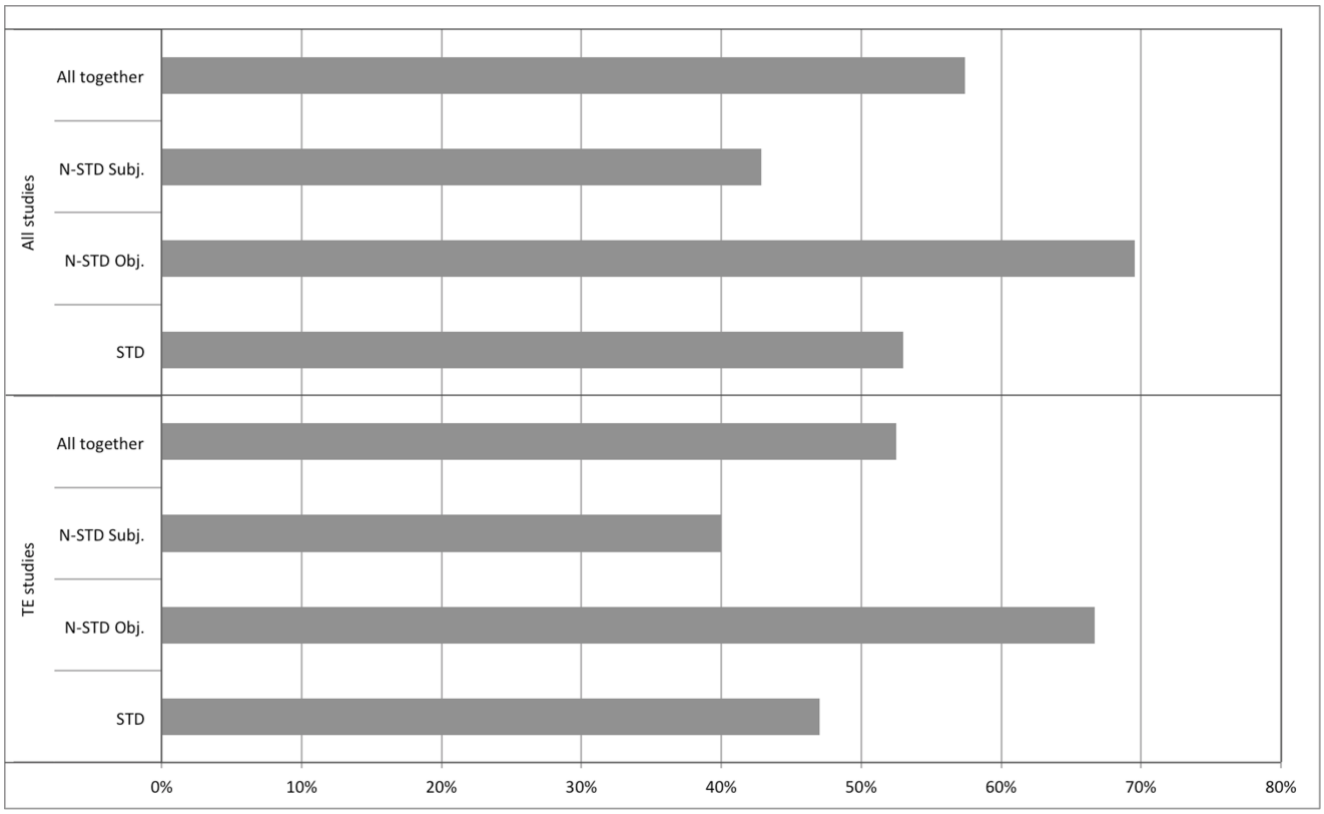
All studies
All together
N-STD Subj.
N-STD Obj.
STD
TE studies
All together
N-STD Subj.
N-STD Obj.
STD
0%
10%
20%
30%
40%
50%
60%
70%
80%

Table 1. PICO criteria and search query related to our literature search.

Patient/Population	Intervention	Comparison	Outcome
Children and adolescents with ASD (0-20 y.o.), both HFA or LFA	TBIs, studies aiming to evaluate TE and/or TU	Compared with typically developed children and/or children with medical condition (ASD or other)	Improvement of school-related skills, evaluation of effectiveness and/or usability
Search query	(autis* OR ASD) AND (mobile device OR tablet OR smartphone OR computer OR technolog*) AND (school* OR pre-school* OR high-school* OR student OR children OR adolescent) AND (intervention OR training) AND (usab* OR effective* OR validat* OR efficacy OR evaluation) AND NOT (gene OR genetic OR protein) AND NOT (brain study OR fMRI study)		

Table 2. Description of **TE studies** (N=22/31). Standardized measurements are shown in bold. (M) means that the measurement was used in a follow-up, and (G) that the measurements was used for assessing the generalization.

Studies	Groups' characteristics			Technology	Targeted skills/behaviors	Intervention settings	Study design	Outcome Measurements	JADAD score
	Treatment group	Control group(s)							
Bartoli, et al. (2014)	5 ASD 6-8 y.o.	5 ASD 6-8 y.o.		*Kinect* Motion-based touchless games	*Cognitive* Selective and sustained attention, visuo-motor integration	Therapeutic center *12 weeks*	*RCT* Pre *vs.* post	**Modified Bell Test** **Cancellation WISC subtest** **Developmental Test of Visual-Motor Integration** Global Weighted Score	1
Bauminger-Zviely, et al. (2013)	14 ASD 10.22 y.o.	8 ASD 9.19 y.o.		*Multitouch table* Join-in and No-Problem software	*Social* Collaboration & conversation	School *12 weeks*	*Crossover study* Pre *vs.* post	Problem-Solving Measure Concept clarification Shared Drawing Task **Happé's Strange Stories (G)**	0
Costescu, Vanderborght & David (2015)	41 ASD 8.4 y.o. (4-13)	40 TD 5.4 y.o. (4-7)		*Robot* Keepon	*Cognitive* Flexibility	Room therapy *Single session*	*Controlled study* Robot *vs.* Human	Number of errors Frequency of shared attention episodes Frequency of positive affects	1
Golan, et al. (2010)	20 (15M; 5F) ASD 5.6 y.o. (4-7)	19 (15M; 4F) ASD 6.2 y.o. (4-8)	18 (12M; 6F) TD 5.4 y.o. (4-7)	*Video DVD* The Transporters	*Emotion* Recognition	Home *4 weeks*	*Quasi-RCT* Pre *vs.* Post	Emotional vocabulary Situation-Facial Expression Matching tasks (G) (3 levels)	2
Gordon, et al. (2014)	17 ASD 10.76 y.o. (6-18)	17 TD 10.94 y.o. (6-18)		*Computer* FaceMaze game	*Emotion* Facial expression production	Laboratory *Single session*	*Controlled study* Pre *vs.* Post	Facial expressions production quality (rated by undergraduate students)	1
Grossman, Peskin & San Juan (2013)	20 (18M; 2F) ASD 8.7 y.o. (7-11)	19 (16M; 3F) ASD 9.6 y.o. (7-13)		*Computer* Gruffee task	*Communication* Communicative clarity	Laboratory *1 week*	*RCT* Pre *vs.* Post	Gruffe tasks (Character and vehicles) Magic tricks task (G) + Follow-up (M)	2
Hopkins, et al. (2011) *LFA groups*	14 (13M; 1F) ASD (LFA) 10.57 y.o.	11 (10M; 1F) ASD (LFA) 10.31 y.o.		*Computer* FaceSay software	*Social* Recognition and interactions	School *6 weeks*	*RCT* Pre *vs.* Post	**Ekman's test** **Benton Facial Recognition test** **Social Skills Rating System (G)** Social interactions observation (G)	2
Hopkins, et al. (2011) H*FA groups*	11 (9M; 2F) ASD (HFA) 9.85 y.o.	13 (12M; 1F) ASD (HFA) 10.05 y.o.							

Table 2. *(continued)*.

Jeong, et al. (2015)	7 (6M; 1F) ASD 10.14 y.o. (6-13)	7 (6M; 1F) ASD 10.29 y.o. (6-13)	*Robot* iRobi	*Emotion* Emotional vocabulary	Not reported *10-20 weeks*	*Controlled study* Pre *vs.* Post Robot *vs.* Computer	Number and diversity of emotional words + Follow-up (M)	0
Lorenzo, et al. (2016)	20 (14M; 6F) ASD 7-12 y.o.	20 (15M; 5F) ASD 7-12 y.o.	*Virtual reality*	*Emotion*	Laboratory *40 weeks*	*RCT* Pre *vs.* Post VR *vs.* Computer	Situation identification (scored by evaluator) Behavioral observations: emotional responses, appropriate behaviors, compliance with the behavior guideline Behavioral data extracted by the system Teachers' interviews (G)	1
Ploog, Banerjee & Brooks (2009)	9 (8M; 1F) ASD 12.9 y.o. (5-18)	9 (7M; 2F) TD 8.0 y.o. (5-11)	*Computer*	*Communication* Prosody	School, Home or Laboratory *Single session*	*Controlled study* Pre *vs.* Post ASD *vs.* TD	Success rate	1
Pop, et al. (2013)	7 ASD 4-9 y.o.	7 ASD 4-9 y.o. 6 ASD 4-9 y.o.	*Robot* Probo	*Social*	Not reported *Single session*	*RCT* Robot *vs.* Computer *vs.* Control	Level of prompting	1
Pop, et al. (2014)	5 ASD 4-7 y.o.	6 ASD 4-7 y.o.	*Robot* Probo	Social Play Engagement	Room therapy *Single session*	*RCT* Pre *vs.* Post	Behavioral observations : Play skills Engagement in play Social skills	2
Rice, et al. (2015)	16 ASD (HFA) 7.68 y.o. (5-11)	15 ASD (HFA) 7.87 y.o. (5-11)	*Computer* FaceSay software	*Emotion* Recognition, mentalizing	School *10 weeks*	*RCT* Pre *vs.* Post	**NEPSY Affect Recognition subtest** **NEPSY ToM subtest** **Social Responsiveness Scale (G)** Social Interaction observations (G)	1
Rodriguez & Cummings (2016)	20 (18M; 2F) ASD and/or SLI 7.4 y.o. (6-10)	11 (8M; 3F) ASD and/or SLI 7.9 y.o. (6-10)	*Tablet (iPad)* Language Builder	*Academic* Language	School *8 weeks*	*Controlled study* Pre *vs.* Post	**Expressive Vocabulary Test-2** **Peabody Picture Vocabulary Test-4** **Clinical Evaluation of Language Fundamentals-4** **Teacher Ratings of Oral Language and Literacy**	0
Salvador, Silver & Mahoor (2015)	11 (9M; 2F) ASD 9.1 y.o. (7-13)	11 (6M; 5F) TD 8.8 y.o. (7-13)	*Robot* Zeno	*Emotion* Recognition	Laboratory *Single session*	*Controlled study* ASD *vs.* TD	Recognition accuracy score	1
Silver & Oakes (2001)	11 ASD 13 y.o. (10-18)	11 ASD 13 y.o. (10-18)	*Computer* EmotionTrainer software	*Emotion* Recognition & prediction	School *2 weeks*	*RCT* Pre *vs.* Post	**Spence's Facial Expression Photographs** **Emotion Recognition Cartoons** Ongoing data from the software **Happé's Strange Stories (G)**	3

Table 2. *(continued)*.

Srinivasan, et al. (2015)	12 (11M; 1F) ASD 7.52 y.o. (5-12)	12 (10M; 2F) ASD 7.88 y.o. (5-12)	12 (11M; 1F) ASD 7.36 y.o. (5-12)	*Robot* Nao	*Repetitive behaviors* *Emotion*	Not reported *8 weeks*	*RCT* Pre *vs.* Post Robot *vs.* Rhythm *vs.* Control	Behavioral observations: Repetitive and maladaptive behaviors Affective states	2
Srinivasan, Eigsti, Gifford, et al. (2016)				*Robot* Nao	*Communication* Spontaneous and Responsive	Not reported *8 weeks*	*RCT* Pre *vs.* Post Robot *vs.* Rhythm *vs.* Control	**Joint Attention Test** Behavioral observations: Social attention patterns	2
Srinivasan, Eigsti, Neelly, et al. (2016)				*Robot* Nao	*Social*	Not reported *8 weeks*	*RCT* Pre *vs.* Post Robot *vs.* Rhythm *vs.* Control	**Joint Attention Test** Behavioral observations: Responses to social bids Vocalization/verbalization patterns	2
Whalen, et al. (2010)	22 ASD 3-6 y.o.	25 ASD 3-6 y.o.		*Computer* TeachTown game	*Social* *Academic* *Cognitive*	School *12 weeks*	*RCT* Pre *vs.* Post	**Expressive Vocabulary Test-2** **Peabody Picture Vocabulary Test-4** **Brigance Inventory of Early Development** Ongoing data from the software	1
Young & Posselt (2012)	13 ASD 4-8 y.o.	12 ASD 4-8 y.o.		*Video DVD* The Transporters	*Emotion* Recognition	Home *3 weeks*	*RCT* Pre *vs.* Post	**NEPSY Affect Recognition subtest** **Faces Task** **Social Communication Questionnaire (G)**	1
Zheng, Young, et al. (2016)	8 ASD 3.83 y.o.	8 TD 3.61 y.o.		*Robot* Nao	*Imitation*	Laboratory *single session*	*Controlled study* Robot *vs.* Human	Attention paid to the administrator Imitation performance	1

Table 3. Description of **TU studies** (N=6/31). Standardized measurements are shown in bold.

Studies	Groups' characteristics		Technology	Targeted skills/behaviors	Intervention settings	Study design	Outcome Measurements	JADAD score
	Treatment group	Control group						
Bekele, et al. (2013)	10 (8M; 2F) ASD (HFA) 14.7 y.o. (13-17)	10 (8M; 2F) TD 14.6 y.o. (13-17)	*Virtual reality*	*Emotion* Facial emotion recognition	Laboratory *Single session*	*Controlled study* ASD *vs.* TD	Performance (success, confidence, latency) Eye-tracking (target, duration, frequency)	0
Bekele, et al. (2014)	6 ASD (HFA) 4.7 y.o. (2-5)	6 TD 4.4 y.o. (2-5)	*Robot* Nao	*Cognitive* Joint attention	Laboratory *Single session*	*Controlled study* Robot *vs.* Human	Gaze to Administrator Level of Prompting Target success Hit frequency	1
Falkmer, et al. (2014)	14 (9M; 5F) ASD or DS 14.1 y.o. (12-16)	23 (14M; 9F) TD 11.6 y.o. (7-15)	*Smartphone* Safeway2school	*Autonomy* School bus safety	Parent's choice *Single session*	*Controlled study* Dis *vs.* TD	Questionnaires on intervention content Questionnaires on trust and acceptance Eye-tracking data (target, duration, frequency)	0
Grynszpan, Martin & Nadel (2008)	10 (10M; 0F) ASD 12.83 y.o.	10 (8M; 2F) TD 9.58 y.o.	*Computer* Game software	*Emotion* Facial emotion recognition	School *12 weeks*	*Controlled study* Pre *vs.* post Rich *vs.* simple interface Real faces *vs.* cartoons	Number of scenarios per session Number of trials per scenario Average duraction of scenarios Number of clicks per utterance Number of correct facial expression recognition	0
Valadão, et al. (2016)	5 (4M; 1F) ASD 7-8 y.o.	5 TD 7-8 y.o.	*Robot* Maria	*Social*	Laboratory *Single session*	*Controlled study* Pre *vs.* Post ASD *vs.* TD	**Goal Attainment Scaling** Lickert-scaled questionnaire Social abilities observations	0
Zheng, Warren, et al. (2016)	8 ASD 2.19 y.o. (0-3)	8 TD 1.33 y.o. (0-3)	*Computer* Intelligent learning environment	*Cognitive* Early social orienting skills	Laboratory *Single session*	*Controlled study* ASD *vs.* TD	Level of prompting Time spent to hit the target	0

Table 4. Description of **TE-TU studies** (N=3/31). Standardized measurements are shown in bold.

Studies	Groups' characteristics			Technology	Targeted skills/behaviors	Intervention settings	Study design	Outcome measurements	JADAD score
	Treatment group	Control group(s)							
Fage (2015)	5 ASD (LFA) 13-16 y.o.	5 TD 13-16 y.o.		*Tablet* Emotomètre	*Emotion* Self-regulation	School *12 weeks*	*Controlled study* Pre *vs.* Post	**EQCA-VS** **USE questionnaire** Questionnaire on child's usage	0
Fage, et al. (2016) *Exp. 1*	5 (5M; 0F) ASD (LFA) 13-16 y.o.	5 (4M; 1F) ASD (LFA) 13-16 y.o.		*Tablet* Activity schedule	*Autonomy* School routines, action planification	School *12 weeks*	*Controlled study* Pre *vs.* Post	Questionnaire on child's usage Questionnaire on performance quality Log data from device	0
Fage, et al. (2016) *Exp. 2*		5 (1M; 4F) ID 13-16 y.o.		*Tablet* Activity schedule	*Autonomy* School routines, action planification	School *12 weeks*	*Controlled study* Pre *vs.* Post	Questionnaire on child's usage Questionnaire on performance quality Log data from device	0
Sitdhisanguan, et al. (2012) *Exp. 1*	4 ASD (LFA) 3-5 y.o.	4 ASD (LFA) 3-5 y.o.	4 ASD (LFA) 3-5 y.o.	*Computer + Tangible Interface*	*Academic* Shape matching	Overtime clinic *1 week*	*Controlled study* Pre *vs.* Post Mouse *vs.* Touch *vs.* Tangible	Number of assists	0
Sitdhisanguan, et al. (2012) *Exp. 2*	8 ASD (LFA) 3-5 y.o.	4 ASD (LFA) 3-5 y.o.	8 ASD (LFA) 3-5 y.o.	*Computer + Tangible Interface*	*Academic* Color recognition	Overtime clinic *4 weeks*	*Controlled study* Pre *vs.* Post Mouse *vs.* Touch *vs.* Tangible	Child's score Time needed to complete the task Learning efficacy score	0

Table 5. Methodological characteristics and results of included studies: recapitulative table. Large effect sizes ($d > 0.7$) are shown in bold.

Study purpose	Study	Technology *Outcome*	Study design		JADAD	Primary Outcomes				Generalization				Results' consistency		Effect Size (SD) *Cohen's d*
						Standardized		*Non-standardized*		*Standardized*		*Non-standardized*				
						Obj.	*Subj.*	*Obj.*	*Subj.*	*Obj.*	*Subj.*	*Obj.*	*Subj.*			
TE	Silver & Oakes (2001)	Computer *Emotion*	RCT	Pre vs. post	3	Y	N	Y	N	Y	N	N	N	Slightly Positive	★★	**0.92 (0.31)**
	Bartoli, et al. (2014)	Kinect *Cognitive*	RCT	Pre vs. post	1	Y	N	Y	N	N	N	N	N	Highly Positive	★★★	**1.30 (1.22)**
	Whalen, et al. (2010)	Computer *Social, Academic & Cognitive*	RCT	Pre vs. Post	1	Y	N	Y	N	N	N	N	N	Limited evidence	★	0.66 (0.45) *Preschool* 0.33 (0.11) *K-1*
	Srinivasan, Eigsti, Neelly, et al. (2016)	Robot *Social*	RCT	Pre vs. post Robot vs. Rhythm vs. Control	2	Y	N	N	Y	N	N	N	N	Limited evidence	★	-0.40 (0.21)
	Srinivasan, Eigsti, Gifford, et al. (2016)	Robot *Communication*	RCT	Pre vs. post Robot vs. Rhythm vs. Control	2	Y	N	N	Y	N	N	N	N	Limited evidence	★	-0.06 (0.38)
	Hopkins, et al. (2011)	Computer *Social*	RCT	Pre vs. post	2	Y	N	N	N	N	Y	N	Y	Slightly Positive	★★	0.22 (0.66) *LFA* 0.14 (0.95) *HFA*
	Rice, et al. (2015)	Computer *Emotion*	RCT	Pre vs. post	1	Y	N	N	N	N	Y	N	Y	Slightly Positive	★★	0.63 (0.68)
	Young & Posselt (2012)	Video DVD *Emotion*	RCT	Pre vs. post	1	Y	N	N	N	N	Y	N	N	Slightly Positive	★★	**0.80 (0.93)**
	Lorenzo, et al. (2016)	Virtual Reality *Emotion*	RCT	Pre vs. post VR vs. Computer	1	N	N	Y	Y	N	N	N	Y	Highly Positive	★★★	**1.30 (0.70)**
	Grossman, Peskin & San Juan (2013)	Computer *Communication*	RCT	Pre vs. post	2	N	N	Y	N	N	N	Y	N	Slightly Positive	★★	0.48 (0.70)
	Pop, et al. (2013)	Robot *Social*	RCT	Robot vs Computer vs Control	2	N	N	Y	N	N	N	N	N	Limited evidence	★	**1.35 (0.38)**
	Golan, et al. (2010)	Video DVD *Emotion*	Quasi-RCT	Pre vs. post	2	N	N	Y	N	N	N	Y	N	Highly Positive	★★★	**1.50 (0.14)**
	Pop, et al. (2014)	Robot *Social & Play*	RCT	Pre vs. post	1	N	N	N	Y	N	N	N	N	Limited evidence	★	**0.80 (0.46)**
	Srinivasan, et al. (2015)	Robot *Emotion & Behavior*	RCT	Pre vs. post Robot vs. Rhythm vs. Control	2	N	N	N	Y	N	N	N	N	Limited evidence	★	-0.16 (0.23)

Table 5. *(continued)*.

TE	Rodriguez & Cummings (2016)	Tablet *Language*	Controlled trial	Pre vs. post	0	Y	Y	N	N	N	N	N	N	Limited evidence	★	0.44 (0.45)
	Bauminger-Zviely, et al. (2013)	Multitouch table *Social*	Controlled trial	Pre vs. post	0	N	N	Y	Y	Y	N	N	N	Highly Positive	★★★	**0.74 (0.21)**
	Jeong, et al. (2015)	Robot *Emotion*	Controlled trial	Pre vs. post	0	N	N	Y	N	N	N	N	N	Highly Positive	★★★	*Not computable*
	Ploog, Banerjee & Brooks (2009)	Computer *Communication*	Controlled trial	Pre vs. post ASD vs. TD	1	N	N	Y	N	N	N	N	N	Highly Positive	★★★	-0.69 (0.86)
	Costescu, Vanderborght & David (2015)	Robot *Cognitive*	Controlled trial	Robot vs. Human	1	N	N	Y	N	N	N	N	N	Slightly Positive	★★	0.57 (0.42)
	Zheng, Young, et al. (2016)	Robot *Imitation*	Controlled trial	Robot vs. Human	1	N	N	Y	N	N	N	N	N	Slightly Positive	★★	0.13 (0.36)
	Salvador, Silver & Mahoor (2015)	Robot *Emotion*	Controlled trial	ASD vs. TD	1	N	N	Y	N	N	N	N	N	Limited evidence	★	-0.07 (0.36)
	Gordon, et al. (2014)	Computer *Emotion*	Controlled trial	Pre vs. post	1	N	N	N	Y	N	N	N	N	Highly Positive	★★★	**0.85 (0.99)**
TU	Valadão, et al. (2016)	Robot *Social*	Controlled trial	ASD vs. TD	0	Y	N	Y	Y	N	N	N	N	Slightly Positive	★★	*Not computable*
	Falkmer, et al. (2014)	Smartphone *Autonomy*	Controlled trial	DIS vs. TD	0	N	N	Y	Y	N	N	N	N	Slightly Positive	★★	-0.16 (0.47)
	Bekele, et al. (2013)	Virtual Reality *Emotion*	Controlled trial	Robot vs. Human	0	N	N	Y	N	N	N	N	N	Highly Positive	★★★	**-0.86 (1.26)**
	Zheng, Warren, et al. (2016)	Computer *Cognitive*	Controlled trial	ASD vs. TD	0	N	N	Y	N	N	N	N	N	Highly Positive	★★★	-0.15 (0.08)
	Grynszpan, Martin & Nadel (2008)	Computer *Communication*	Controlled trial	Pre vs. post Rich vs. simple interface Real faces vs. cartoons	0	N	N	Y	N	N	N	N	N	Slightly Positive	★★	0.12 (0.47)
	Bekele, et al. (2014)	Robot *Cognitive*	Controlled trial	ASD vs. TD	1	N	N	Y	N	N	N	N	N	Slightly Positive	★★	**1.31 (1.15)**
TE+TU	Fage (2015)	Tablet *Emotion*	Controlled trial	Pre vs. post	0	N	Y	N	Y	N	N	N	N	Highly Positive	★★★	*Not computable*
	Fage, et al. (2016)	Tablet *Autonomy*	Controlled trial	Pre vs. post	0	N	N	Y	Y	N	N	N	N	Highly Positive	★★★	**1.00 (0.32)**
	Sitdhisanguan, et al. (2012)	Computer *Academic*	Controlled trial	Pre vs. post Mouse vs. Touch vs. Tangible	0	N	N	Y	N	N	N	N	N	Highly Positive	★★★	**2.05 (3.18)**

Appendix

SIGN ratings (SIGN, 2008). The eight ratings are as follow: 1++: High quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias. 1+: Well-conducted meta-analyses, systematic reviews, or RCTs with a low risk of bias. 1-: Meta-analyses, systematic reviews, or RCTs with a high risk of bias. 2++: High quality systematic reviews of case control or cohort or studies or high quality case control or cohort studies with a very low risk of confounding or bias and a high probability that the relationship is causal. 2+: Well-conducted case control or cohort studies with a low risk of confounding or bias and a moderate probability that the relationship is causal. 2-: Case control or cohort studies with a high risk of confounding or bias and a significant risk that the relationship is not causal. 3: Non-analytic studies, such as case reports or case series. 4: Expert opinion.

JADAD score (Jadad, et al., 1996). The Jadad score is computed from three criteria: randomization, double-blind assessment and dropout/exclusion report. One point is given to the study if (a) there is a randomization for the allocation of groups, (b) the study was conducted with a double-blind assessment, and (c) the authors explicitly reported the number of participants that were excluded and/or who have abandoned, and the reasons why. The points for randomization and double blind are given only if there is a statement to evoke it in the paper. Also, if there were no dropout and no abandon, an explicit statement have to be provided in order to allocate the point. An additional point is given if (a) the method of randomization is described and appropriate, and/or (b) the method of double blind is described and appropriate. Conversely, a point is removed if (a) the method of randomization is described and inappropriate, and (b) the method of double blind is described and inappropriate.