Virtual environments as an assessment modality with pediatric ASD populations: a brief report

Tyler C. Duffield, Thomas D. Parsons, Allisen Landry, Shaza Karam, Tiffany Otero, Sarah Mastel and Trevor A. Hall

ABSTRACT
Virtual environments (VEs) have demonstrated promise as a neuropsychological assessment modality and may be well suited for the evaluation of children suspected of having an autism spectrum disorder (ASD). Some recent studies indicate their potential for enhancing reliability, ecologically validity, and sensitivity over traditional neuropsychological evaluation measures. Although research using VEs with ASD is increasing to the degree that several reviews of the literature have been conducted, the reviews to date lack rigor and are not necessarily specific to cognitive or neuropsychological assessment as many focus on intervention. The aim of this project was to comprehensively examine the current literature status of neuropsychological assessment in pediatric ASD using VEs by conducting a systematic review. Specifically, psychometric comparisons of VEs to traditional neuropsychological assessment measures that examined reliability, validity, and/or diagnostic accuracy for pediatric individuals, age 18 and below, with ASD were sought. The search using key words yielded 899 manuscripts, 894 of which were discarded for not meeting inclusion criteria. The remaining five met exclusion criteria. Therefore, the systematic review was modified to a brief report. These findings (or lack thereof) indicate a significant gap in the literature in that psychometric comparisons of these tools for the neuropsychological assessment of pediatric individuals with ASD are lacking. An important future direction of research will be extending the demonstrated incremental validity of VE neuropsychological assessment with other neurodevelopmental (e.g., attention-deficit/hyperactivity disorder) and adult populations to pediatric ASD populations.

A concern that has been raised about neuropsychological assessments is their limited ecological validity (Burgess et al., 2006; Chaytor & Schmitter-Edgecombe, 2003; Chaytor, Schmitter-Edgecombe, & Burr, 2006). For youth, this often relates to the ability to predict classroom performances. Due to being conducted in sterile (i.e., quiet and generally distraction free) settings, traditional paper-and-pencil testing batteries do not reflect...
real-world executive function demands and/or distractions (Parsons, 2016b) and were developed to assess cognitive "constructs" (e.g., memory) without regard for their ability to predict the functional capacities of youth for successful performance in everyday activities (Burgess et al., 2006). Hence, predicting real-world functioning with any rigor, as is often requested of neuropsychology referrals, can be difficult.

This may be especially true for the neuropsychological assessment of higher functioning children with autism spectrum disorder (ASD) where deficits in "cold" and "hot" executive functions can equally impact classroom performance, which comprises constant cognitive and social demands (see Parsons, 2014 for discussion of the social orienting hypothesis). "Cold" executive functions refer to mechanistic higher order cognitive abilities, such as working memory. "Hot" executive functions refer to cognitive abilities supported by emotional awareness and social perception in a real-world environment. Historically, "cold" executive function impairments have been viewed as a primary deficit in ASD (Ozonoff, Pennington, & Rogers, 1991), but these impairments disappear when these same abilities are measured using analog-computerized versions of tests (e.g., Ozonoff, 1995). Thus, removal of the social interaction with the tester during traditional measurement, or hot executive demands, resulted in commensurate performance between typically developing individuals and individuals with ASD. Our understanding of the overlap between "cold" and "hot" executive functions, such as working memory processes support of emotion recognition and social inference skills during dynamic social interactions (Zimmerman, Ownsworth, O’Donovan, Roberts, & Gullo, 2016), continues to grow. Given that current paper-and-pencil neuropsychological measures can demonstrate impaired cold executive function that actually reflects hot executive function deficits (e.g., Ozonoff, 1995), there is need for measures of both cold and hot executive functions that capture expression of real-world deficits. This has important implications for the predictive value and generalizability of assessment data for everyday functioning of youth with ASD in simulations of real-world activities, such as those performed in the classroom, compared to current assessment methods.

Virtual environments (VEs) have demonstrated increased reliability, validity, and ecological validity for assessing cognitive functions in neuropsychological assessment (Parsons, Carlew, Magtoto, & Stonecipher, 2017). For example, response to naturalistic distracter conditions differentiated ADHD children (Parsons, Bowerly, Buckwalter, & Rizzo, 2007) and ASD adults (Parsons & Carlew, 2016) from controls on assessments delivered via the virtual classroom environment when traditional neuropsychological measures could not. Even more studies have been conducted to examine the utility of VEs for intervention and training of skills for persons with ASD. Parsons (2016a) identified six VE-specific reviews for ASD populations that generally focused on the effectiveness (e.g., use, application, understanding, and interpretation) of VE interventions (e.g., behavioral and social skills studies), each suggesting that VEs have a valuable potential to improve clinical and real-world outcomes (e.g., gainful employment) for individuals with ASD. The demonstrated utility of VEs for neuropsychological assessment combined with the enhanced ecological validity offered by VEs has led to a growing body of literature suggesting VEs may be ideal for the neuropsychological assessment of ASD (Parsons & Carlew, 2016).

Clearly, research interest in the use of VEs in ASD is increasing. However, if the suggestion that VEs may be ideal for the neuropsychological assessment of ASD is to be
further substantiated, several gaps in the existing literature should be addressed. Nolin et al. (2016) note that a large majority of neuropsychological studies employing VEs have examined adults, such as Parsons and Carlew’s (2016) comparison of only adult neurotypical and individuals with ASD across modalities of the Stroop task. It is now well established that findings in adult populations cannot be inferred down to apply to pediatric populations (e.g., Giza, Mink, & Madikians, 2007). Further, and of equal importance, the psychometric comparability of various presentation modalities (i.e., paper-and-pencil, computerized, and VEs) needs to be established for use in pediatric ASD populations. A joint position paper by the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology has stressed that “new and different tests” require new psychometric studies and must provide users with core information regarding test reliability, validity, accuracy, and utility (Bauer et al., 2012). An aim of the current review was to examine the status of research employing psychometric comparisons of traditional neuropsychological assessment modalities (i.e., pencil-and-paper or analogous computerized versions) to VEs in pediatric ASD populations.

Method

Guidelines for conducting a systematic review discussed by Khan, Kunz, Kleijnen, and Antes (2003) and Uman (2011) were followed. The “free-form” question was as follows: Do VEs perform equal-to-or-better-than traditional modalities in the neuropsychological assessment of ASD? The aim was not to examine diagnostic assessments of ASD. The “structured” question formula (intervention/exposure for population with condition) was as follows: The outcome of interest was psychometric comparisons of VEs to traditional measures that examined reliability, validity, and/or diagnostic accuracy (sensitivity, specificity, positive predictive power, and negative predictive power) for pediatric individuals (18 and below) with ASD. Research designs that were quantitative comparisons were sought.

The first and last author formulated a comprehensive list of search terms. A search strategy biased toward sensitivity (retrieving a high proportion of relevant studies) rather than specificity (retrieving a low proportion of irrelevant studies) was employed due to concern of a small number of available articles (Uman, 2011). Individual search terms were examined within the PsycINFO and PubMed databases in 9/2016. A comprehensive list of search terms was established through individual term search outcomes, performed by the first author, demonstrating a high numeric yield of articles found. High-yield search terms were those terms that encompassed similar search term results. For example, a search for “virtual” also includes results of a search for “virtual reality.” High-yield search terms were also unique terms that did not have multiple iterations to consider (e.g., pediatric). Application of these criteria resulted in retention of seven condition terms (i.e., autism, Asperger, Asperger’s, Asperger’s syndrome, Asperger syndrome, neurodevelopmental, pervasive developmental disorder), four population terms (i.e., child, pediatric, adolescent, teen), and eight exposure terms (i.e., virtual, immersive, immersion, simulation, game, gaming, artificial, augmented). Next, searches were conducted within the PsycINFO and PubMed databases, with no date restrictions, for every combination of condition, population, and intervention/exposure terms (e.g., autism and child and virtual) by a different author in 10/2016.
Inclusion and exclusion criteria were established based upon first and last author consensus. Inclusion criteria included (1) quantitative comparison of VEs to traditional assessment measure/paradigm (i.e., paper-and-pencil test/questionnaire or analogous computerized version of test) for emotional, behavioral, or cognitive variables; (2) diagnosis of ASD; (3) English-language journal; and (4) peer-reviewed journal. Exclusion criteria included (1) qualitative comparisons (i.e., no empirical comparison); (2) participants greater than 18 years of age; (3) dissertations; (4) reviews or meta-analyses; (5) articles that focus on treatment/intervention; and (6) augmented reality or artificial intelligence. The search terms “artificial” and “augmented” were included even though augmented reality and artificial intelligence were exclusion criteria again in the service of sensitivity. Different authors/and or organizations use different terms, particularly concerning mixed reality (i.e., hybrid of natural world and virtual graphics/assets).

Two authors then independently evaluated abstracts of each article to determine whether they met criteria for inclusion, followed by full text review to assess if criteria were met for exclusion in this review. An interrater reliability analysis using the Kappa statistic was performed to determine consistency between both authors who reviewed abstracts. A data-extraction template and design-based quality assessment template were modified from previous public domain templates.

Results and discussion

Please refer to Figure 1 for the flow diagram of the systematic review. Every combination of condition, population, and exposure terms resulted in 2544 manuscripts. A total of 1645 duplicates were removed (i.e., sensitivity bias) with 899 remaining manuscripts. A total of 894 manuscripts were discarded, as they did not meet inclusion criteria (entirely due to a lack of quantitative comparison of VEs to traditional neuropsychological assessment measure/paradigm). The interrater reliability for the raters was found to be kappa = 0.75 (p < .001), 95% CI (0.412, 1.086). This constitutes a substantial level of agreement (Landis & Koch, 1977).

Given the limited number of articles to survive application of inclusion and exclusion criteria, full-text review included even discrepant articles from both authors who reviewed abstracts (three discrepancies as one reviewer retained two abstracts and one reviewer retained five abstracts). Of the five remaining manuscripts, all five were excluded due to meeting exclusion criteria. A data-extraction template and design-based quality assessment template were not applied as no manuscripts survived inclusion and exclusion criteria. A qualitative analysis of manuscripts demonstrated considerable overlap with articles discussed in Parsons (2016a) conceptual review.

Psychometric comparisons between paper-and-pencil tests and novel technological translations of traditional tests are vital in understanding the comparability of assessment results across modalities (Bauer et al., 2012). In the same way that test revisions (e.g., WAIS-IV/WMS-IV have been criticized for failing to consider back-compatibility issues that may invalidate clinical interpretations) using traditional print publishing can have unintended consequences (Loring & Bauer, 2010), novel technological presentation modalities (e.g., VEs) must be validated to ensure that they retain equivalent psychometric properties. Comparability must be empirically demonstrated. This is ever more relevant as more researchers and various clinical specialties expand their use of testing into other technological formats (De Marco & Broshek, 2016).
While a number of studies have been completed to validate VEs for neuropsychological assessment in general (see Parsons et al., 2017 for review), psychometric studies aimed at validating VEs for pediatric assessment of ASD specifically appear largely the exception rather than the rule. However, there are some psychometric comparisons being made. In pediatric populations, for example, the continuous performance task in a virtual classroom has been examined for typically developing (Iriarte et al., 2016; Nolin et al., 2016) and other neurodevelopmental disorders (Adams, Finn, Moes, Flannery, & Rizzo, 2009; Bioulac et al., 2012; Diaz-Orueta et al, 2013; Gilboa, Rosenblum, Fattal-Valevski, Toledano-Alhadeh, & Josman, 2011; Neguț, Jurma, & David, 2017; Nolin, Stipanicic, Henry, Joyal, & Allain, 2012; Parsons et al., 2007, 2010; Pollak et al., 2009). In addition, a virtual Stroop task has been examined in adult ASD populations (Parsons & Carlew, 2016).

Psychometric comparisons of these tools for the neuropsychological assessment of pediatric individuals with ASD are lacking as the current review demonstrated, although the use of VEs in ASD populations, primarily for intervention purposes, continues to expand (Wass & Porayska-Pomsta, 2014). This is a particularly important area of future
research considering most identification, and thus testing, treatment, and training occur in childhood for ASD. An important future research agenda will be to attempt to replicate the demonstrated incremental validity of virtual assessment environments with other neuro-developmental and adult populations with pediatric ASD populations. This is especially evident considering assessment environments have the capability to also function as treatment environments, which has been noted to have a valuable potential to improve clinical and real-world outcomes for individuals with ASD (e.g., Parsons, 2016a).

**Limitations**

A major limitation of this review is that the quality and bias of the status of VE assessment in pediatric ASD populations were not assessed, which is a primary aim of systematic reviews. It can be argued that given the emphasis on sensitivity (and thus inclusivity), the quality of virtual assessment tool comparisons in pediatric ASD simply cannot be assessed due to a lack of studies engaging in such work. A large number of exclusion criteria may also be a limitation, but these authors chose exclusion criteria that are common to thorough systematic reviews conducted in the available literature. Additionally, not including non-English language articles may have resulted in missed psychometric comparisons of virtual assessment tools in pediatric ASD as many international groups are conducting VE research. Again though, not including non-English language articles is common practice given that international articles available may not be translated from native languages, or potentially untranslated (not having their sense expressed in another language).

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**ORCID**

Tyler C. Duffield 💬 [http://orcid.org/0000-0003-3101-7027](http://orcid.org/0000-0003-3101-7027)

**References**


