Immersive Virtual Reality Implementations in Developmental Psychology

Supplemental Analyses

Araiza-Alba, Keane, Beaudry, Kaufman

30 October 2020

Overview

In this document, we report the key findings from the papers we discussed in our manuscript. Many of the empirical papers did not report complete statistical analyses or effect sizes; therefore, when possible, we calculated these using the information provided in the original papers. The data and code for these analyses are available on the Open Science Framework at: https://osf.io/f96hc/

Interpretation of effect sizes

The effect sizes we report can be interpreted according to the following standards:

- Cohen's *f* for one-way ANOVAs: small = 0.10, medium = 0.25, large = 0.40
- Cohen's *d* for comparing two means: small = 0.20, medium = 0.50, large = 0.80
- Cohen's *h* for proportions: small = 0.20, medium = 0.50, large = 0.80
- partial eta-squared (n_p^2) for ANOVAs and MANOVAs: small = .01, medium = .06, large = .14 (Cohen, 1998)
- Vargha-Delaney's A for Mann-Whitney U tests produces a value between 0 and 1. When A is exactly 0.5, the two techniques achieve equal performance; when A is less than 0.5, the first technique is worse; when A is more than 0.5, the second technique is worse. The closer to 0.5, the smaller the difference between the techniques. Specifically, for A, a small effect = .56, medium effect = .64, and large effect = .71 (Vargha & Delaney, 2000).

IVET used as pain distraction for children

Dahlquist et al. (2009)

Dahlquist et al. (2009) ran a 2 (Age: younger, older) X 3 (Condition: baseline, distraction-only, distraction+helmet[VR]) mixed-factorial ANOVA on participants' pain tolerance scores (no transformation reported).

Their ANOVA revealed a significant main effect of condition, $F(2,78) = 7.84$, $p < .001$, $f = .62$, with a large effect size. They did not report the means or any post hoc comparisons for this effect.

They also reported a significant Age X Condition interaction on pain tolerance, $F(2,78) = 3.82$, $p < .05$, $f = .43$, with a large effect size. The only post hoc comparison they reported (with enough detail for reanalysis) was comparing the pain tolerance scores of younger and older children in the distraction+helmet condition.

We used the R package {BSDA} (Arnholt & Evans, 2017) to compute the t-test statistic based on the summary statistics reported in the paper, and we used the R package {compute.es} (Re, 2013) to calculate the effect size (Cohen's *d*) with 95% confidence intervals (CI). Within the distraction+helmet (VR) condition, older children showed significantly higher pain tolerance ($M = 70.08$, $SD = 71.22$) than the younger children ($M = 31.74$, $SD = 40.36$), $t = 2.19$, $p = .034$, $d = 0.66$ [0.08, 1.24].

Dahlquist et al. (2010)

Dahlquist et al. (2010) used one-way within-subjects ANOVA to compare the three conditions (baseline, distraction-only, distraction+helmet [VR]) on log-transformed pain tolerance scores. They reported that their ANOVA showed a significant main effect of condition, $F(2,98) = 20.45$, $p < .001$, with a large effect size ($f = 0.65$). They did not report effect sizes for the follow-up paired-sample t-tests (with Bonferroni correction).

We could not recreate their post hoc analyses from the summary data, so we report the t-values & p-values from Dahlquist et al. (2010, p. 622). Then, we used the R package {compute.es} (Re, 2013) to calculate the associated effect sizes (Cohen's *d*) with 95% confidence intervals for the three comparisons.

These posthoc tests showed that children's pain tolerance was significantly lower in the baseline condition ($M = 1.25$, $SD = 0.23$) than in the traditional distraction condition ($M = 1.45$, $SD = 0.33$), $t(49) = 6.01$, $p < .001$, $d = 0.70$ [0.30, 1.11], and the VR distraction condition ($M = 1.44$, $SD = 0.38$), $t(49) = 4.68$, $p < .001$, $d = 0.60$ [0.20, 1.01]. There was no significant difference between the two distraction conditions, $t(49) = 0.11$, $p > .91$, $d = 0.03$ [-0.36, 0.42].

Hua et al. (2015)

Hua et al. (2015) reported only the p-values from their statistical comparisons between the VR distraction and standard distraction groups on each of the 10 dependent variables. For each of these comparisons, we used the summary statistics (means & standard deviations) reported in the paper to compute the inferential statistics and effect sizes. For two of these comparisons (noted in Table S1. 1), our calculations produced slightly different p-values than those reported in the paper, but the conclusions were unchanged.

Specifically, we used the R package {BSDA} (Arnholt & Evans, 2017) to conduct independent t-tests and the R package {compute.es} (Re, 2013) to calculate the effect sizes (Cohen's *d*) and accompanying 95% confidence intervals [reported in square brackets].

Table S1. 1

Hua et al. (2015)

Scale	Measure	VR Distraction M (SD)	Standard Distraction M (SD)	t-value	p-value	Cohen's d [95% CI]
FACES	self-reported pain before dressing change	0.85 (1.12)	1.63 (1.39)	2.49	0.016	0.62 [0.12, 1.12]
	self-reported pain during dressing change	2.42 (1.85)	4.19 (2.12)	3.58	0.001	0.89 [0.38, 1.40]
	self-reported pain after dressing change	2.48 (1.80)	3.38 (1.48)	2.2	0.031†	0.55 [0.05, 1.04]
VAS	observed pain before dressing change (caregivers)	0.99 (0.68)	1.87 (2.14)	2.22	0.033†	0.56 [0.06, 1.05]
	observed pain during dressing change (caregivers)	4.35 (2.64)	6.25 (2.84)	2.79	0.007	0.69 [0.19, 1.19]
	observed pain after dressing change (caregivers)	2.67 (1.89)	5.94 (1.59)	7.56	< .001	1.87 [1.29, 2.45]
FLACC	observed distress before dressing change (nurses)	0.52 (0.69)	0.84 (0.75)	1.79	0.079	0.44 [-0.05, 0.94]
	observed distress during dressing change (nurses)	4.18 (2.97)	7.36 (3.47)	3.96	< .001	0.99 [0.47, 1.50]
	observed distress after dressing change (nurses)	3.68 (2.73)	5.79 (3.84)	2.55	0.014	0.64 [0.14, 1.13]
	duration of dressing changes	22.3 (7.85)	27.9 (6.83)	3.07	0.003	0.76 [0.26, 1.26]

Note: IVR group had 33 participants and the control group had 32 participants. † computed p-values differ from the original paper, but the conclusions are the same.

Jeffs et al. (2014)

Jeffs et al. (2014) compared three groups (standard care, *n* = 10; passive distraction, *n* = 10; and virtual reality, *n* = 8) on reported pain. The findings supporting our descriptive summary are available on page 402 of their paper. To compare the procedural pain reported by participants in the three groups, they used a linear regression model that adjusted for age, sex, state anxiety, opioid analgesic use, treatment length, and pre-procedural pain. From this model, they estimated participants' procedural pain scores.

We were unable to reproduce these analyses because the paper did not include the relevant details. Specifically, we could not compare the three groups in their procedural pain scores because, although Jeffs et al. (2014) reported the estimated means, they did not report a measure of variance or indicate what the error bars in Figure 2 represent. Furthermore, although Jeffs et al. (2014) did report effect sizes comparing the groups (ranging from 0.535 to 1.25), they did not specify the measure of effect size; thus, we cannot interpret the size of these effects.

Kipping et al. (2012)

Kipping et al. (2012) compared pain-related scores for patients in the IVR distraction group (*n* = 20) and the standard distraction group (*n* = 21).

For the continuous measures, they analyzed mean change scores by subtracting the baseline measures from the removal and application scores for the outcome variables (self-reported pain, self-reported nausea, nurses' pain ratings, caregivers' pain ratings). For these analyses, they reported the means and standard deviations, as well as the p-values for the independent sample t-tests (but not the associated t-values).

For these measures, we used the R package {BSDA} (Arnholt & Evans, 2017) to calculate the t-values from the summary statistics and we used the R package {compute.es} (Re, 2013) to compute effect sizes (Cohen's *d*) with 95% confidence intervals [in square brackets]. See Table S1. 2for the summary and inferential statistics. Our calculated p-values were slightly different than those reported in the original paper (for all but one analysis: nurses' pain ratings during removal), but the conclusions were the same.

Unlike the other measures, the rescue doses outcome was binary data. Of the 20 adolescents in the VR group, only 3 needed rescue doses (15%). Of the 21 adolescents in the SD group, 9 needed rescue doses (43%). The paper reports that they ran a chi-square test, which produced a significant result at p = .05.

Using the base R package {stats} (R Core Team, 2020), we tried to reproduce this analysis, but our chi-square test revealed a non-significant difference, $\chi^2(1) = 2.61$, p = .106. A one-sided z-test of proportions produced a similar result to that of the original paper, z = 1.62, p = 0.053; however, it is worth noting that this p-value still exceeded the alpha level of .05. Using the R package {pwr} (Champely, 2020), we calculated Cohen's h, a measure of effect size for proportions. This analysis revealed a medium-large effect size (h = 0.63 [0.02, 1.24]) with a large 95% confidence interval. Taken together, these results suggest this finding might not be reliable.

Table S1. 2

Kipping et al. (2012)

Scale	Measure	IVR Distraction	Standard Distraction	t-value	p-value	Cohen's d
		M(SD)	M(SD)	(df = 39)		
VAS	self-reported pain (removal)	2.9 (2.3)	4.2 (3.2)	1.49	0.145†	0.46 [-0.16, 1.09]
	self-reported pain (application)	2.33 (3.4)	3.8 (3.6)	1.34	0.187†	0.42 [-0.20, 1.04]
	self-reported nausea (removal)	-0.7 (1.1)	-0.3 (1.5)	0.97	0.338†	0.30 [-0.31, 0.92]
	self-reported nausea (application)	-0.3 (1.0)	-0.5 (1.3)	-0.55	0.585†	-0.17 [-0.79, 0.44]
FLACC	nurses' pain ratings (removal)	2.9 (2.4)	4.7 (2.5)	2.35	0.024	0.73 [0.10, 1.37]
	nurses' pain ratings (application)	1.9 (2.8)	3.0 (2.8)	1.26	0.216†	0.39 [-0.23, 1.01]
VAS	caregivers' pain ratings (removal)	3.5 (2.5)	3.8 (3.2)	0.33	0.741†	0.10 [-0.51, 0.72]
	caregivers' pain ratings (application)	2.6 (3.5)	2.2 (4.0)	-0.34	0.736†	-0.11 [-0.72, 0.51]

Note: IVR group: *n* = 20; standard distraction group: *n* = 21. † computed p-values differ from the original paper, but the conclusions are the same.

Sil et al. (2012)

Sil et al. (2012) used a one-way within-subjects ANOVA to compare the three conditions (baseline, traditional distraction, VR distraction) on children's (n = 62) log-transformed pain tolerance scores. They reported that their ANOVA showed a significant main effect of condition, F(1,122) = 19.15, p < .001, with a large effect size (f) of 0.56. They did not report effect sizes for the follow-up paired-sample t-tests (with Bonferroni correction).

We could not recreate their analyses from the summary data, so we report the t-values & p-values from Sil et al. (2012, p. 6). Then, we used the R package {compute.es} (Re, 2013) to calculate the associated effect sizes (Cohen's d) with 95% confidence intervals.

The posthoc tests showed that children's pain tolerance was significantly lower in the baseline condition (M = 1.37, SD = 0.28) than in the traditional distraction condition (M = 1.57, SD = 0.45), t(61) = 4.99, p < .001, d = 0.53 [0.18, 0.89], and the VR distraction condition (M = 1.56, SD = 0.44), t(61) = 5.44, p < .001, d = 0.52 [0.16, 0.87]. There was no significant difference between the two distraction conditions, t(61) = 0.34, p = .73, d = 0.02 [-0.33, 0.37].

IVET used as a neuropsychological tool for children

Diaz-Orueta et al. (2014)

In our opinion, relatively little additional analysis of the findings of Diaz-Orueta et al. (2014) was necessary because they reported their analyses in appropriate detail. We include relevant statistical details about the convergent validity of the AULA Nesporala and Conners CPT in our manuscript. The only additional analysis needed was to calculate the effect size associated with the Mann-Whitney U analyses reported on p. 338 of their paper. We calculated Vargha-Delaney's A (using formula 7 from Ruscio, 2008). Unfortunately, we could not calculate confidence intervals without the raw data.

Given the large number of comparisons reported in the paper, we report only the smallest and largest effect sizes between the AULA scores comparing children undergoing treatment (n = 29) and those without treatment (n = 28); all comparisons were significant. The smallest effect was found for RT commissions without distractors. Children undergoing treatment responded faster, M = 667.87 (SD = 240.43), than those without treatment, M = 795.31 (SD = 292.8), U = 282.5, z = -1.971, p = 0.049, A = .65. The largest effect between the AULA scores was for X task motor activity. Children undergoing treatment scored lower, M = 0.86 (0.84), than those without treatment, M = 1.7 (1.36), U = 203.5, z = -3.233, p = 0.001, A = .75.

Readers are encouraged to refer to the original paper for further details about the other measures.

Nolin et al. (2016)

Like Diaz-Orueta et al. (2014), in our opinion, relatively little additional analysis of the findings of Nolin et al. (2016) was necessary. We include relevant statistical details about the convergent validity of the ClinicaVR and VIGIL-CPT, as well as the test–retest reliability of the ClinicaVR in our manuscript.

Negut et al. (2016)

They used a mixed factorial design with two between-subject factors: test condition (VC assessment vs. traditional CPT) and clinical status (children with ADHD [*n* = 33] vs. typically-developing children [*n* = 42]), and one within-subjects factor: test modality (with vs. without distractors). They ran a MANCOVA with age and IQ as covariates. We report their V-, *F*-, and *p*-values for the MANCOVA results, but we also used the R package {effectsize} (Ben-Shachar, Makowski, & Lüdecke, 2020) to calculate partial eta squared (n_p^2) with 90% confidence intervals for these effects.

We also used the R package {BSDA} (Arnholt & Evans, 2017) to recalculate the post hoc follow-up tests and the R package {compute.es} (Re, 2013) to compute the effect size (Cohen's *d*) with 95% confidence intervals (CI). See **Table S2. 1** for relevant descriptive and inferential statistics for the post hoc tests for commission errors, omission errors, and total correct responses. We could not report the post hoc test for the response time variable because were unable to obtain the relevant summary statistics from the original paper.

For our purposes, two key findings emerged from their multivariate analysis:

1. A significant effect of clinical status showing that typically-developing children performed better than children with an ADHD diagnosis, V = .30, *F*(4,66) = 7.06, p < .001, n_p^2 = 0.300, 90% CI [0.127, 0.417].

2. No significant interaction between test condition and clinical status of the children, V = .08, *F*(4,66) = 1.60, p > .05, n_p^2 = 0.088, 90% CI [0.000, 0.173], which suggests that the children with ADHD and the typically-developing children performed similarly regardless of the type of test used.

Table S2. 1

Negut et al. (2016)

Measure	ADHD group M (SD)	Control group M (SD)	*t*-value	*p*-value	Cohen's d [CI]
commission errors	30.48 (28.26)	10.92 (12.95)	3.68	0.001	0.93 [0.45, 1.41]
omission errors	39.12 (45.63)	18.9 (22.95)	2.32	0.025	0.58 [0.12, 1.05]
total correct responses	79.6 (24.61)	107.61 (28.23)	-4.58	< .001	-1.05 [-1.53, -0.56]

Note: ADHD group: *n* = 33; control group: *n* = 42. † computed t-values differ from the original paper, but the conclusions are the same.

Pollak et al. (2009)

Pollak et al. (2009) conducted 3 (Test: TOVA, No VR-CPT, VR-CPT) X 2 (Group: ADHD vs. control) mixed-factorial ANOVAs on each of the DVs: reaction time, variability of reaction time, errors of omission, errors of commission. We report their F values for the ANOVA results, but we used the base R stats package (R Core Team, 2020) to recalculate the *p*-values and the R package {effectsize} (Ben-Shachar et al., 2020) to calculate partial eta squared (n_p^2) with 90% confidence intervals for these effects.

We also used the R package {BSDA} (Arnholt & Evans, 2017) to compute the post hoc follow-up tests and the R package {compute.es} (Re, 2013) to compute the effect size (Cohen's *d*) with 95% confidence intervals. See Table S2. 2 for relevant descriptive and inferential statistics for the post hoc tests comparing the groups (ADHD vs. control) on each of the tests.

Note 1: Pollak et al. (2009) report "approximate Cohen's *d* values" with no further explanation. Given the lack of additional information, we are not concerned about the consistent, but minimal, differences between our computed values and those they reported.

Note 2: We believe their Table 1 has a typo (p. 5). For the control group's TOVA score on the errors of omission measure, they report a mean of .29 and a standard deviation of 48. We presume they missed the decimal place in the standard deviation value, so we have corrected it in our table. However, despite computing a smaller *d* value for this comparison than that reported in the original paper (*d* = 1.01 vs. 1.32, respectively), our conclusion was different than theirs. Specifically, we found a significant difference between the two groups on errors of omission in the TOVA condition, but they reported (in text) that this difference was not significant. As such, future research is needed to shed light on this result.

They reported a significant Test X Group interaction on response time, $F(2,68) = 3.8$, $p = 0.03$, $n_p^2 = 0.101$, 90% CI [0.007, 0.213], and errors of omission, $F(2,68) = 9.9$, $p <.001$, $n_p^2 = 0.226$, 90% CI [0.085, 0.354]. They did not find significant effects on variability of response time and errors of commission; they did not report the associated statistics for these results.

In addition to the significant interaction effects, group had a significant main effect on all four measures: reaction time, $F(1,34) = 4.6$, $p = 0.04$, $n_p^2 = 0.119$, 90% CI [0.003, 0.301]; variability of reaction time, $F(1,34) = 5.6$, $p = 0.02$, $n_p^2 = 0.141$, 90% CI [0.011, 0.327]; errors of omission, $F(1,34) = 28.8$, $p < .001$, $n_p^2 = 0.459$, 90% CI [0.251, 0.61]; and errors of commission, $F(1,34) = 16.0$, $p < .001$, $n_p^2 = 0.320$, 90% CI [0.119, 0.497].

Table S2. 2

Pollak et al. (2009)

Measure	Test	Control group M (SD)	ADHD group M (SD)	*t*-value	*p*-value	Cohen's *d* [CI]
reaction time	TOVA	393 (65)	451 (114)	1.94	0.062	0.61 [-0.05, 1.27]
	No VR-CPT	578 (89)	609 (178)	0.68	0.499	0.21 [-0.43, 0.86]
	VR-CPT	546 (83)	677 (142)	3.48	0.001	1.10 [0.41, 1.80]
variability of reaction time	TOVA	115 (39)	154 (61)	2.35	0.025	0.75 [0.08, 1.42]
	No VR-CPT	121 (49)	150 (51)	1.76	0.087	0.58 [-0.08, 1.24]
	VR-CPT	128 (26)	145 (46)	1.41	0.169	0.45 [-0.21, 1.10]
errors of omission	TOVA	0.29 (0.48)	3.05 (3.69)	3.31	0.004	1.01 [0.32, 1.69]
	No VR-CPT	2.65 (1.58)	11.75 (9.69)	4.14	0.001	1.26 [0.55, 1.97]
	VR-CPT	5.06 (5.1)	22.34 (14.67)	4.93	< 0.001	1.52 [0.79, 2.26]
errors of commission	TOVA	2.63 (3.15)	5.85 (4.03)	2.73	0.01	0.88 [0.20, 1.56]
	No VR-CPT	0.56 (0.58)	1.86 (1.11)	4.56	< 0.001	1.43 [0.71, 2.16]
	VR-CPT	0.94 (0.84)	2.37 (1.91)	3.02	0.005	0.94 [0.26, 1.62]

Note: ADHD group: *n* = 20; control group: *n* = 17. The computed d-values are different than those in Pollak et al. (2009), but the conclusions remain the same.

Rodriguez et al. (2018)

Participants (control: n = 101; ADHD: n = 237) were randomly assigned to complete either completed the AULA Nesplora (VR) test or the TOVA test. We cannot directly compare the means across the two tests because the values are interpreted differently (high scores on VR are indicative of deficit; high scores on the TOVA are indicative of good executive functioning).

Rodriguez et al. (2018) report two key analyses that relevant to our manuscript:

1. They conducted two MANCOVAs (one for VR & one for TOVA) to examine differences across the four groups (control & 3 ADHD presentation groups: Inattentive, Impulsive and Hyperactivity, Combined) on the dependent variables (omissions, commissions, response time, and variability) across the two halves of the tests. We report their lambdas, *F*-values, and *p*-values to supplement the information in the manuscript, and we used the R package {effectsize} (Ben-Shachar et al., 2020) to calculate partial eta squared (n_p^2) with 90% confidence intervals for these effects.

- The MANCOVA for VR, with age as a covariate, was statistically significant, λ = .506, $F(24,429) = 4.93$, $p < .001$, $n_p^2 = 0.216$, with significant univariate effects for all dependent variables (.057 ≥ n_p^2s ≤ .228).

- The MANCOVA for TOVA, with IQ as a covariate, was not statically significant, λ = .050, $F(24,465) = 1.18$, $p = .249$, $n_p^2 = 0.057$.

2. Rodriguez et al. (2018) also explored whether the dependent variables (omissions, commissions, response time, and variability; split by test half) provided by the two tests could correctly predict group membership. Both analyses were significant, but the discriminant function for the VR test correctly classified more of the sample (56.60%) than the function for the TOVA (33.70%).

Taken together, these findings suggest that the AULA Nesplora was able to discriminate between children with and without ADHD symptoms, whereas the TOVA was not able to discriminate between the two groups.

Bioulac et al. (2012)

Bioulac et al. (2012) used a 2 (Group: ADHD vs. control) X 2 (Test: VR classroom vs. CPT) mixed-factorial design, with type of test manipulated within subjects.

They used non-parametric tests (namely, Mann-Whitney U tests) to compare the two groups' performance on each test. They did not report enough data to allow us to compute an effect size measure for this comparison. As such, in Table S2. 3, we provide the summary statistics and the *p*-values reported in the original paper (p. 517).

Table S2. 3

Bioulac et al. (2012) Group Comparisons for Each Test

Test	Measure	Control group M (SD)	ADHD group M (SD)	*p*-value
VC	correct hits	85.81 (8.48)	67.95 (11.54)	< .001
	commissions	14.87 (10.05)	21.05 (9.87)	< .05
	correct hits reaction time	0.54 (0.08)	0.52 (0.1)	ns
	reaction time variability	0.2 (0.05)	0.21 (0.06)	ns
	commissions reaction time	569 (203.24)	582 (139.88)	ns
CPT	correct hits	312.16 (8.94)	297.2 (12.89)	< .001
	commissions	22.56 (6.35)	26.25 (5.06)	ns
	correct hits reaction time	412.31 (63.35)	460.15 (86.15)	< .05
	reaction time standard deviation	8.08 (2.15)	14.47 (5.13)	< .001

Note: Control group: $n = 16$; ADHD group: $n = 20$.

According to Bioulac et al. (2012), the deterioration of performance across blocks—measured by correct hits—for children with ADHD was significant for the Virtual Classroom test, χ^2(df = 4) = 25,299, $p < .001$, but not for the CPT test, χ^2 (df = 5) = 10,145, $p = .07$. For each test, they reported Mann-Whitney U values comparing the correct hits of the two groups in each block (with the exception of Blocks 1 and 2 for the CPT test). To compute an effect size for these comparisons, we used the reported Mann-Whitney U values (p. 517–518) to compute Vargha-Delaney's A (using formula 7 from Ruscio, 2008). Unfortunately, we could not calculate confidence intervals around this effect size without the raw data. Table S2. 4 provides the Mann-Whitney U-values and associated *p*-values reported in the original paper along with our computed A values.

Table S2. 4

Bioulac et al. (2012) Correct Hits Across Blocks For Each Test

Virtual Classroom				Continuous Performance Test			
Block	U-value	*p*-value	A-value	Block	U-value	*p*-value	A-value
1	51	< .001	0.84	1	--‡	--	--
2	80.5	< .05	0.75	2	--‡	--	--
3	27	< .001	0.92	3	41	< .001	0.87
4	55.5	< .001	0.83	4	71	< .01	0.78
5	66.5	< .01	0.79	5	57	< .001	0.82
				6	83.5	< .05	0.74

Note: Control group: *n* = 16; ADHD group: *n* = 20; the VC test had only 5 blocks; ‡ no U-value reported for block.

Gilboa et al. (2015)

Gilboa et al. (2015) had participants with an acquired brain injury (ABI; *n* = 41) and those without an ABI (*n* = 35). Participants completed the TEA-Ch (completed only by those with ABI) and the Virtual Classroom assessment, and their parents completed the CPRS-R:S. We do not report the results of the TEA-Ch because only children with ABI completed the test; thus, we cannot compare their performance against the control group of children without ABI.

Gilboa et al. (2015) used univariate ANOVAs to compare the two groups on the measures from the VC and CPRS-R:S tests. In Table S2. 5, we report their *F*-values from the ANOVA results (p. 5), but we used the base R package {stats} (R Core Team, 2020) to recalculate the *p*-values and the R package {effectsize} (Ben-Shachar et al., 2020) to calculate Cohen's *d* with 95% confidence intervals. We followed their decision to correct for multiple comparisons (.05 / 4 = .0125).

Table S2. 5

Gilboa et al. (2015)

Test	Measure	Control group M (SD)	ABI group M (SD)	*F*-value	*p*-value	Cohen's *d* [CI]
VC	total correct hits	88.4 (10.8)	76.3 (17.19)	13.09	0.0005	-0.83 [-1.30, -0.36]
	errors of commission	7.7 (10.3)	13.4 (21.6)	2.22	0.1404	0.33 [-0.13, 0.78]
	reaction time	49.8 (14.9)	50.6 (11.3)	0.009	0.9247	0.06 [-0.39, 0.51]
	head movement	195.2 (125.3)	200.9 (135.1)	0.26	0.6116	0.04 [-0.41, 0.49]
CPRS-R:S	opposition	54.8 (11.7)	58.8 (12.9)	1.93	0.1690	0.32 [-0.14, 0.78]
	inattention	51.3 (9.4)	56.2 (10.7)	4.27	0.0423	0.48 [0.02, 0.95]
	hyperactivity	52.1 (8.5)	58.5 (13.9)	5.38	0.0232	0.55 [0.08, 1.01]
	index ADHD	51.0 (8.4)	56.6 (11.0)	6.04	0.0164†	0.57 [0.10, 1.03]

Note: For the VC test, ABI group: *n* = 41 and control group: *n* = 35; one participant in each group did not complete the CPRS-R:S test. As such, the error degrees of freedom were 75 for the VC measures and 73 for the CPRS-R:S measures. † Reported as significant at .01 in Gilboa et al. (2015), but it is not significant at the corrected alpha (.0125) based on our computed *p*-values.

IVET used as a social-skills training tool for children with autism spectrum disorder

Ip et al. (2018)

Ip et al.'s (2018) participants were children with an ASD diagnosis (*n* = 72). Half were assigned to the VR-enabled training group and half to the control group. Participants completed pre- and post-training assessments on a range of measures, including affective expressions (i.e. emotion expression and regulation), social reciprocity, and emotion recognition (Faces test & Eyes test).

Ip et al. (2018) reported a significant Group X Time interaction on affective expressions, $F(1,70) = 5.22$, $p = .025$, $n_p^2 = .069$, and social reciprocity, $F(1,70) = 7.769$, $p = .007$, $n_p^2 = .100$. The Group X Time interaction was not significant on either measure of emotion recognition: Faces test, $F(1,70) = 0.188$, $p = .666$, $n_p^2 = .003$: Eyes test, $F(1,70) = 0.470$, $p = .495$, $n_p^2 = .007$.

In Table S3. 1, we report the summary statistics and the *d*-, *t*-, and *p*-values reported in Table 3 of Ip et al. (2018) comparing the pre- and post-training scores within each training group. We did not recalculate the effect size measure because our calculations suggest that they (appropriately) reported d_{av} for these correlated means (Lakens, 2013); unfortunately, we could not calculate confidence intervals around this parameter without the raw data.

To determine if the post-training scores in the two groups were significantly different for the two measures that had significant interactions (affective expressions and social interaction), we used the R package {BSDA} (Arnholt & Evans, 2017) to conduct independent t-tests. We also used the R package {compute.es} (Re, 2013) to calculate the effect sizes (Cohen's *d*) and accompanying 95% confidence intervals for these independent means.

For affective expressions, the VR group's post-training score was significantly higher than the control group's post-training score, $t(58.36) = 3.788$, $p = 0.0004$, $d = 0.89$ [0.41, 1.38]. For social reciprocity, the VR group's post-training score was also significantly higher than the control group's post-training score, $t(63.62) = 2.815$, $p = 0.0065$, $d = 0.66$ [0.19, 1.14].

Table S3. 1

Ip et al. (2018)

Measure	VR training					Control group				
	Pre-training M (SD)	Post-training M (SD)	Cohen's *d*	*t*-value	*p*-value	Pre-training M (SD)	Post-training M (SD)	Cohen's *d*	*t*-value	*p*-value
affective expressions	18.9 (3.57)	20.2 (3.00)	0.39	-2.174	0.037	17.0 (4.01)	16.6 (4.85)	0.09	0.909	0.37
social reciprocity	20.2 (3.43)	21.8 (2.99)	0.50	-3.987	< .001	19.6 (4.14)	19.4 (4.15)	0.05	0.293	0.771
Faces test	16.6 (2.36)	17.2 (2.33)	0.26	-1.233	0.226	15.4 (3.01)	15.7 (2.9)	0.1	-0.721	0.476
Eyes test	12.8 (3.54)	13.5 (3.98)	0.19	-1.37	0.178	12.1 (3.61)	12.3 (3.7)	0.06	-0.33	0.745

Note: Control group: n = 36; experimental (VR training) group: n = 36

Lorenzo et al. (2016)

In Lorenzo et al. (2016), 40 children with an ASD diagnosis completed 40 sessions over a 10-month period. Half of the children completed the training in a semi-CAVE system (referred to as the experimental group) and half used a non-immersive VR application (referred to as the control group).

The presentation of the results made it difficult to reproduce and/or summarize the analyses; however, we present the results that were accessible. The summary statistics reported in the paper were not statistically analyzed, as far as we could tell (see Table S3. 2). As such, we used the R package {BSDA} (Arnholt & Evans, 2017) to conduct independent t-tests comparing the two groups and the R package {compute.es} (Re, 2013) to calculate the effect sizes (Cohen's *d*) and accompanying 95% confidence intervals.

Table S3. 2

Lorenzo et al. (2016)

Measure	Session	Experimental group M (SD)	Control group M (SD)	*t*-value	*p*-value	Cohen's *d* [CI]
identifying the social situation	Average	5.2 (0.4)	4.5 (0.14)	7.39	< .001	2.34 [1.53, 3.14]
adequate behaviors	1	8.5 (4.1)	7.6 (4.5)	0.66	0.513	0.21 [-0.41, 0.83]
adequate behaviors	4	14.0 (2.9)	10.5 (3.0)	3.75	0.001	1.19 [0.51, 1.86]
inadequate behaviors	Average	16.3 (6.5)	22.2 (4.1)	-3.43	0.002	-1.09 [-1.75, -0.42]
teachers' scores	Average	4.0 (1.0)	3.4 (0.4)	2.49	0.02	0.79 [0.14, 1.43]

Note: Both groups had 20 participants; Average = The score was the aggregate score across sessions

Reproducible Code Statement

I used R (Version 3.6.3; R Core Team, 2020) and the R-packages *apa* (Version 0.3.3; Gromer, 2020; Aust & Barth, 2018), *BSDA* (Version 1.2.1; Arnholt & Evans, 2017), *compute.es* (Version 0.2.5; Re, 2013), *dplyr* (Version 1.0.2; Wickham et al., 2020), *effectsize* (Version 0.3.3; Ben-Shachar et al., 2020), *effsize* (Version 0.8.0; Torchiano, 2020), *finalfit* (Version 1.0.2; Harrison, Drake, & Ots, 2020), *fmsb* (Version 0.7.0; Nakazawa, 2019), *forcats* (Version 0.5.0; Wickham, 2020), *ggplot2* (Version 3.3.2; Wickham, 2016), *here* (Version 0.1; Müller, 2017), *kableExtra* (Version 1.1.0; Zhu, 2019), *knitr* (Version 1.29; Xie, 2015), *lattice* (Version 0.20.41; Sarkar, 2008), *lavaan* (Version 0.6.7; Rosseel, 2012), *lme4* (Version 1.1.23; Bates, Mächler, Bolker, & Walker, 2015), *MASS* (Version 7.3.52; Venables & Ripley, 2002), *Matrix* (Version 1.2.18; Bates & Maechler, 2019), *papaja* (Version 0.1.0.9842; Aust & Barth, 2018), *PearsonDS* (Version 1.1; Becker & Klößner, 2017), *purrr* (Version 0.3.4; Henry & Wickham, 2020), *pwr* (Version 1.3.0; Champely, 2020), *readr* (Version 1.3.1; Wickham, Hester, & Francois, 2018), *shiny* (Version 1.5.0; Chang, Cheng, Allaire, Xie, & McPherson, 2020), *stringr* (Version 1.4.0; Wickham, 2019), *tibble* (Version 3.0.3; Müller & Wickham, 2020), *tidyr* (Version 1.1.1; Wickham & Henry, 2020), *tidyverse* (Version 1.3.0; Wickham, Averick, et al., 2019), *tinytex* (Version 0.25; Xie, 2019), and *WebPower* (Version 0.5.2; Zhang & Mai, 2018) for all analyses.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology: Cognition, 4*, 1–12. https://doi.org/10.338/fpsyg.2013.00863

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods, 13*(1), 19–-30. https://doi.org/10.1037/1082-989X.13.1.19

Vargha, A., & Delaney, H. (2000). A critique and improvement of the "CL" Common Language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25*(2), 101–132. https://doi.org/10.2307/1165329

R package References

Arnholt, A. T., & Evans, B. (2017). *BSDA: Basic statistics and data analysis*.

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bates, D., & Maechler, M. (2019). *Matrix: Sparse and dense matrix classes and methods*. Retrieved from https://CRAN.R-project.org/package=Matrix

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Becker, M., & Klößner, S. (2017). *PearsonDS: Pearson distribution system*. Retrieved from https://CRAN.R-project.org/package=PearsonDS

Ben-Shachar, M. S., Makowski, D., & Lüdecke, D. (2020). Compute and interpret indices of effect size. *CRAN*. Retrieved from https://github.com/easystats/effectsize

Champely, S. (2020). *Pwr: Basic functions for power analysis*. Retrieved from https://CRAN.R-project.org/package=pwr

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2020). *Shiny: Web application framework for r*. Retrieved from https://CRAN.R-project.org/package=shiny

Gromer, D. (2020). *Apa: Format outputs of statistical tests according to apa guidelines*. Retrieved from https://CRAN.R-project.org/package=apa

Harrison, E., Drake, T., & Ots, R. (2020). *Finalfit: Quickly create elegant regression results tables and plots when modelling*. Retrieved from https://CRAN.R-project.org/package=finalfit

Henry, L., & Wickham, H. (2020). *Purrr: Functional programming tools*. Retrieved from https://CRAN.R-project.org/package=purrr

Müller, K. (2017). *Here: A simpler way to find your files*. Retrieved from https://CRAN.R-project.org/package=here

Müller, K., & Wickham, H. (2020). *Tibble: Simple data frames*. Retrieved from https://CRAN.R-project.org/package=tibble

Nakazawa, M. (2019). *Fmsb: Functions for medical statistics book with some demographic data*. Retrieved from https://CRAN.R-project.org/package=fmsb

R Core Team. (2020). *R: A language and environment for statistical computing*. Retrieved from https://www.R-project.org/

Re, A. C. D. (2013). Compute.es: Compute effect sizes. In *R Package*. Retrieved from https://cran.r-project.org/package=compute.es

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/

Sarkar, D. (2008). *Lattice: Multivariate data visualization with r*. Retrieved from http://lmdvr.r-forge.r-project.org

Torchiano, M. (2020). *Effsize: Efficient effect size computation*. doi: 10.5281/zenodo.1480624

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth). Retrieved from http://www.stats.ox.ac.uk/pub/MASS4/

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H. (2020). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from https://CRAN.R-project.org/package=forcats

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi: 10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2020). *Tidyr: Tidy messy data*. Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., Hester, J., & Francois, R. (2018). *Readr: Read rectangular text data*. Retrieved from https://CRAN.R-project.org/package=readr

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Retrieved from https://yihui.org/knitr/

Xie, Y. (2019). TinyTeX: A lightweight, cross-platform, and easy-to-maintain latex distribution based on tex live. *TUGboat*, (1), 30–32. Retrieved from http://tug.org/TUGboat/Contents/contents40-1.html

Zhang, Z., & Mai, Y. (2018). *WebPower: Basic and advanced statistical power analysis*. Retrieved from https://CRAN.R-project.org/package=WebPower

Zhu, H. (2019). *KableExtra: Construct complex table with 'kable' and pipe syntax*. Retrieved from https://CRAN.R-project.org/package=kableExtra