Supplementary Online Content

Voss C, Schwartz J, Daniels J, et al. Effect of wearable digital intervention for improving socialization in children with autism spectrum disorder: a randomized clinical trial. *JAMA Pediatr*. Published online March 25, 2019. doi:10.1001/jamapediatrics.2019.0285

This supplementary material has been provided by the authors to give readers additional information about their work.

eMethods 1. Study protocol, additional outcome measures, and emotion classifier accuracy measurements

Study Protocol

After determining eligibility, all participants attended their first appointment ("intake"), where they consented to participate, completed the study measures described below, and were randomly assigned with a 1:1 ratio to either the treatment or control condition after all baseline measures were completed. Families in the control condition received no study intervention from the intake appointment and continued ABA therapy as usual for six weeks, with an option to cross over to the treatment condition at Post-test 1.

Families in the treatment condition received a thorough demonstration of SG at the end of their intake appointment and were sent home with the SG kit consisting of a "how to use your device" guide, a pair of Google Glasses, a pin-protected android phone with the study app installed, and chargers and cases for both. Each family was asked to use the device at least three times per week at home for 20 minutes each, and once per week for 20 minutes with their BI. Participants were instructed to run each of the three engagement activities at least once, but otherwise chose activities freely. App usage was automatically logged for device usage analysis after participants returned their device. All participants completed a second appointment after intake, Post-test 1, during which the control cohort crossed over to the treatment condition, received the device demonstration and took home the device kit after completing study measures. During the post-test 1 appointment, the treatment cohort returned their device and completed a semi-structured interview on their experience in addition to all study measures. All participants then attended a third appointment, Post-test 2, during which the control cohort returned their device and completed a semi-structured interview and study measures, while the treatment cohort completed all final study measures. Only the control cohort attended a final fourth appointment to match the treatment condition arm (Post-test 3), during which the participants completed all study measures. Figure 2 shows the consort flow diagram by study appointment.

Additional Outcome Measures

We used an additional tool, the *Brief Observation of Social and Communicative Change (BOSCC[1,2])*, as a potential secondary outcome measure and a more child-focused endpoint. The BOSCC is a structured play-based lab assessment intended to measure change in core symptoms of autism in children. Our intention was to use it as a structured lab assessment to measure social behavior change in our treatment cohort when children were not wearing the glasses. The assessment was recorded using both a Go-Pro camera in a corner of the room and glasses worn by the blinded clinical coordinator –simply serving as a second camera.

While the framework for administration of BOSCC has been published, the scoring algorithm and validation of use in independent replicate samples has not yet been published. Therefore, we were unable to use the BOSCC as initially planned. We intend to explore the data collected for this measure in the future if/when the algorithms emerge in the literature. We also hope to develop automated tools for this analysis in the near future.

Emotion Classifier Accuracy Measurements

We measured accuracy in three ways:

First, we performed general accuracy validation for methods used in the machine learning system that powers the Superpower Glass application against the Extended Cohn-Kanade Dataset.[3] The facial expression recognition system achieved 97.0% accuracy across over 9000 static images of humans emoting the 8 different emotions embedded in our tool.

Second, during in-lab training and prior to taking the device home for the treatment period, each family member present was asked to express each of the 8 expressions for the unblinded clinical research coordinator, who wore the wearable device. The device was determined to have "failed" if there existed an expression that (1) the device consistently misclassified, but (2) the acting parent could consistently make in a way that the research coordinator could readily identify. These failure criteria were used to judge whether the family would qualify for a participant-specific model as pursued in our uncontrolled design pilot, and we saw no failure of this type for any of the treatment cohort included in the present study.

Third, we focused the empirical assessment of emotion classification accuracy on Guess the Emotion game play sessions in which the parent selects an emotion to start a session, acts the emotion shortly after making the selection and then records the child's guess, whether correct or incorrect. We sampled between two different emotion prompts, prompt A (e.g. "happy") and prompt B (e.g. "sad") selected in turn by the parent. We considered a true positive found an emotion classification prediction of prompt A before prompt B was selected, we called this a positive. With this procedure, we achieved an overall accuracy of 0.72. Although lower than the that seen with the Cohn-Kanade Dataset (which contains manually annotated expressions acted out by adults facing high resolution cameras together with egocentric framing of the head/face), this accuracy is high for active, real use systems like the SG tool. Moreover, due to the heuristic nature of this measurement (due to the computational tractability and reliability of manual measurement of the thousands of images generated per family during use), this value is likely to be an underestimate. In accord with this accuracy assessment, families did not report problems with the emotion classification accuracy during the study.

eMethods 2. Treatment Crossover Analysis

Treatment Crossover Analysis
To take advantage of the additional data available for treatment cohorts, we repeated the exploratory treatment cohort analysis with the *full treatment cohort* including participants who crossed over from the control group. The data was normalized such that the control participants' post-test 1 visit is now labeled as their treatment intake appointment.

Full Treatment Cohort Demographics
We broke the full cohort down by the same analysis groups. None of the group demographics had statistically significant differences.

Cohort	All participants (ITT)			Completers		
	N	Mean Age (SD)	% Male (N)	N	Mean Age (SD)	% Male (N)
Treatment	52	8.33 years (2.06)	90% (N=47)	39	8.49 years (2.20)	90% (N=35)
Control	31	7.94 years (1.69)	84% (N=26)	25	8.44 years (1.83)	84% (N=21)
Total	83	8.18 years (1.93)	88% (N=73)	64	8.47 years (2.05)	88% (N=56)

eTable 1. Primary analysis participant demographics, broken down by analysis groups

Each demographic cohort is representative of each other. None of the group demographics had statistically significant differences.

Cohort	All participants (ITT)			Completers		
	N	Mean Age (SD)	% Male (N)	N	Mean Age (SD)	% Male (N)
Treatment	40	8.64 years (2.52)	93% (N=37)	27	8.76 years (2.86)	93% (N=25)
Control	31	7.94 years (1.69)	84% (N=26)	25	8.44 years (1.83)	84% (N=21)
Total	71	8.38 years (2.46)	89% (N=63)	52	8.45 years (2.06)	90% (N=47)

eTable 2. Primary Moderator Analyses for ABIQ, Age and Gender

Each table presents results for an augmentation of eqn. 1 to include the modifier (e.g., age) and interaction terms with treatment, week, as well as a three-way interaction term with treatment and week. Likelihood ratio (LR) tests then assessed whether the modifier significantly impacted the treatment effect.

ABIQ Moderator Analysis

Measure	**Analysis Cohort**	**Moderator Coefficient** (gender × treatment × time)	**LR Test *p* value**
SRS-2	All Participants (ITT)	0.004	0.218
	Completers	0.003	0.236
EGG	All Participants (ITT)	-0.005	0.109
	Completers	-0.005	0.075
VABS-II Socialization	All Participants (ITT)	0.001	0.538
	Completers	0.011	0.369
NEPSY-II Affect	All Participants (ITT)	0.010	0.103
	Completers	0.010	0.095
CBCL	All Participants (ITT)	0.007	0.070
	Completers	0.008	0.147
VABS	All Participants (ITT)	-0.001	0.552
	Completers	0.001	0.910

Age Moderator Analysis

Measure	**Analysis Cohort**	**Moderator Coefficient** (age × treatment × time)	**LR Test *p* value**
SRS-2	All Participants (ITT)	0.183	0.411
	Completers	0.169	0.446
EGG	All Participants (ITT)	-0.112	0.791

	Completers	-0.107	0.615
VABS-II Socialization	All Participants (ITT)	-0.052	0.797
	Completers	-0.007	0.372
NEPSY-II Affect	All Participants (ITT)	0.028	0.835
	Completers	0.036	0.690
CBCL	All Participants (ITT)	0.029	0.854
	Completers	0.027	0.939
VABS-II	All Participants (ITT)	-0.319	**0.040***
	Completers	-0.283	0.073

Gender Moderator Analysis

Measure	**Analysis Cohort**	**Moderator Coefficient** (gender × treatment × time)	**LR Test *p* value**
SRS-2	All Participants (ITT)	2.10	0.185
	Completers	1.91	0.175
EGG	All Participants (ITT)	2.48	**0.004***
	Completers	2.71	**0.002***
VABS-II Socialization	All Participants (ITT)	-1.68	0.245
	Completers	-1.85	0.189
NEPSY-II Affect	All Participants (ITT)	-0.058	0.986
	Completers	0.035	0.769
CBCL	All Participants (ITT)	0.396	0.568
	Completers	0.320	0.537
VABS-II	All Participants (ITT)	-1.87	0.275
	Completers	-1.75	0.226

eTable 3. Six-week follow-up analysis in treatment-first group for exploratory cohort (N=52) and completers (N=39)

Results from the follow-up analysis conducted on all treatment participants, including the normalized treatment portions of control participants who crossed over, are similar to those of the treatment-first cohort. We caution that cross-over participants start their treatment intake after a first control intake appointment, so potential practice effects may have created a higher baseline for these participants.

Measure	Analysis Cohort	Regression Coefficients		γ_2 Wald Test *p* value
		γ_1	γ_2	
SRS-2	All treatment	-0.337	-0.245	**0.001***
	Treatment Completers	-0.329	0.143	**0.002***
EGG	All treatment	0.770	0.406	**<0.001***
	Treatment Completers	0.741	0.396	**<0.001***
VABS-II Socialization	All treatment	0.330	0.228	**0.024***
	Treatment Completers	0.403	0.248	**0.017***
NEPSY-II Affect	All treatment	0.040	0.020	0.523
	Treatment Completers	0.047	0.054	0.453

eTable 4. Mean change in secondary exploratory outcome measures by cohort subgrouping from intake to post-test 1 using the same mixed effects model

Measure	Analysis Cohort	Mean change from intake to 6-week post-test 1 (SD)	
		Treatment	Control
VABS-II Overall	All participants (ITT)	-0.451 (1.443)	0.889 (1.080)
	Completers	-0.836 (1.499)	0.972 (1.090)
CBCL Overall	All participants (ITT)	-1.130 (1.824)	0.462 (1.606)
	Completers	-1.325 (1.837)	0.820 (1.621)

eResults

Treatment Cohort Usage Statistics

The mean intake to post-test 1 time period for the exploratory cohort was 6.84 weeks (SD=1.73 weeks). Families in the treatment group who returned their device for a conclusion appointment ran at least one recorded session on an average of 11.43 (SD = 5.56) days (48% of the requested dosage of 24 days of device activity). Participants chose to play Guess the Emotion for 40.7% of the sessions, Capture the Smile for 24.0% of the sessions, and Free Play for 35.3% of the sessions. Families ran an average of 3.7 (SD=3.56) sessions with their BI, 62% of the recommended dose.

eReferences

1. Grzadzinski R, Carr T, Colombi C, et al. Measuring Changes in Social Communication Behaviors: Preliminary Development of the Brief Observation of Social Communication Change (BOSCC). *J Autism Dev Disord* 2016; **46**(7): 2464-79.
2. Kitzerow J, Teufel K, Wilker C, Freitag CM. Using the brief observation of social communication change (BOSCC) to measure autism-specific development. *Autism Research* 2016; **9**(9): 940-50.
3. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on; 2010: IEEE; 2010. p. 94-101.