RUNNING HEAD: THOUGHT BUBBLE TRAINING

Brief Report: Further Evidence on the Benefits of Thought-Bubble Training for Theory
of Mind Development in Children with Autism Spectrum Disorders

Jessica Paynter¹ & Candida C. Peterson^{2*}

¹AEIOU Foundation, Brisbane, Australia, ²The University of Queensland, Brisbane,
Australia

*Candida Peterson is corresponding author

Address : Professor C.C. Peterson

School of Psychology, University of Queensland

Brisbane, Queensland 4072 Australia

Phone : +617 3365 6332

Facsimile: +617 3365 4466

Email: candi@psy.uq.edu.au

Note this is the author's copy of the final accepted manuscript, for the published version please see:

Paynter, J., & Peterson, C. C. (2013). Further evidence of benefits of thought-bubble training for theory of mind development in children with autism spectrum disorders. *Research in Autism Spectrum Disorders*, 7(2), 344-348. doi: <http://dx.doi.org/10.1016/j.rasd.2012.10.001>

Abstract

Children with autism spectrum disorders (ASD) routinely fail false belief tests of theory of mind (ToM), even at advanced chronological and mental ages. Initial efforts to train them to pass were largely disappointing. However, a recent study using pictorial thought bubbles showed some promise. With a sample of 24 children with ASD (mean age 7.00 years), we extended its methodology via (a) a non-intervention control group (b) a ToM Scale to examine generalization to non-false-belief ToM concepts and (c) comprehensive assessment of linguistic and nonverbal intelligence. Results revealed even stronger evidence for the benefits of thought-bubble training, including generalization to a broad ToM Scale. Control children showed no significant gains of any kind despite their close match to trained children at pre-test. Theoretical and practical implications of the findings are discussed.

Abbreviations: Autism Spectrum Disorder (ASD); Theory of Mind (ToM); Peabody Picture Vocabulary Test (PPVT)

Keywords: Autism Spectrum Disorders, Theory-of-Mind, Training Studies

*Brief Report: Further Evidence on the Benefits of Thought-Bubble Training for
Theory of Mind Development in Children with Autism Spectrum Disorders*

Theory of mind (ToM) describes the child's understanding of mental states (e.g., knowledge, beliefs) and how they influence behavior (Wellman & Liu, 2004). A key measure is the inferential false belief task requiring predictions about the actions (e.g., search strategies) of protagonists with false beliefs (e.g., about displaced objects' locations). Typically-developing children routinely pass these tests by age 5 (Wellman, Cross & Watson, 2001) but severe delays often accompany a diagnosis of autism (Happé, 1995). Training children with autism to understand false beliefs has produced mixed and uncertain results. However some success has recently been achieved when training incorporates pictorial representations of thinking (e.g., McGregor, Whiten, & Blackburn, 1998; Swettenham, Baron-Cohen, Gomez, & Walsh, 1996; Wellman, Baron-Cohen, Caswell, Gomez, Swettenham, Toye, et al., 2002). Initially, picture-in-the-head training (e.g. McGregor, et al., 1998; Swettenham, et al., 1996) demonstrated only modest success. Pass rates often improved on the exact task that was trained, but typically showed little or no generalisation either to new kinds of false belief tests or even to the same one with different stimuli, in line with a "hacking" (Happé, 1995) interpretation. In other words, rather than gaining any genuine understanding of mental states, children with ASD who are shown to improve via training may instead have simply acquired non-ToM-related heuristics (e.g., "say the wrong answer") that are task-specific. While a more recent study (Wellman et al., 2002) showed greater promise by using the cartoon device of thought bubbles to depict protagonists' true and false beliefs, evidence of generalization to new, untrained false belief concepts remained limited.

There are unresolved questions about Wellman et al.'s pioneering research. Methodologically, perhaps the most serious limitation was lack of a non-intervention control group. Control children with ASD can sometimes improve upon their own pretest false-belief

scores simply via repeated testing and elapsed time, as was highlighted by Fisher and Happé's (2005) study of training in executive functioning. Using pre- and post- false belief tests similar to Wellman et al.'s (2002) across a similar 5-10-day training, Fisher and Happé found an untrained control group gained significantly on a deceptive container false-belief task, even without intervention of any kind. Thus, without a control group, it is unclear whether Wellman et al.'s findings for thought bubbles may have reflected a similar pattern. Furthermore, it is important, both theoretically and practically, to see whether any genuine benefits that conceivably may accrue via the acquisition in skills for using thought bubbles to represent false beliefs are able to generalize even beyond this one isolated ToM concept to other conceptual steps in broader ToM mastery, as assessed by a well-established, psychometrically-reliable ToM Scale (Wellman & Liu, 2004).

In summary, we tested the replicability, validity and robustness of Wellman et al.'s (2002) promising earlier report that systematic training in the use of thought-bubbles with a fresh sample of 17 children aged 4.67 to 12.25 years ($SD = 1.95$) with ASD by including a closely matched control group and a delayed follow-up test. We also assessed gains and generalization more broadly than they had via (a) a larger battery of false belief tasks, and (b) a developmentally-sequential ToM Scale of five steps leading up to and beyond, false belief.

Method

Participants

Twenty-four Australian children independently diagnosed with an autism spectrum disorder (ASD) and enrolled in special classes for autism took part. Their ages ranged from 4.67 to 12.25 years (mean = 7.00, $SD = 1.95$) and all but 3 were male. At pretest, the majority (15/24: 63%) failed both trials of a two-trial changed-location false belief test (Baron-Cohen, Leslie & Frith, 1985) and all (100%) scored below ceiling of a five-task ToM Scale (Wellman & Liu, 2004). The 17 children who were allocated to the training group on an essentially random basis (i.e., (a) timing within the school year and (b) preferences by school staff for greater versus

lesser involvement in research that did not differentially reflect qualities of the children or the educational programs they were being exposed to) did not differ significantly from the 7 included in the control group on any pretest ToM measure (see Table 1). Groups were also closely matched in age, semantic language skill (receptive vocabulary), syntactic ability, and nonverbal intelligence.

Materials and Procedure

Training Phase. Using two-dimensional cardboard stimuli (e.g., dolls, bubble cut-outs, and rooms with door flaps), each child in the experimental group individually received training in how to represent beliefs via cartoon bubbles. We closely followed Wellman et al.'s (2002) exact procedure to create our own scripted demonstrations and questions (available from first author). As per the original, five graduated training stages were encompassed: (a) thought bubbles depict thinking; (b) people think about things they cannot see; (c) when objects change thoughts about them remain constant unless one sees the change; (d) thought bubbles direct searches for hidden things; and (e) obsolete thought bubbles will direct futile searches if objects move. We also added a sixth step: (f) different people have different thought bubbles, depending on what they last saw. To progress to each subsequent stage, children had to reach whatever pre-determined success criterion Wellman et al. specified. Training was complete, and the immediate post-test was given, when all six stages were mastered.

Pretests: Language ability. The Peabody Picture Vocabulary Test (PPVT: Dunn & Dunn, 1997), a standardized test of semantic language skill (receptive vocabulary) with picture pointing responses, was scored via the manual to give a verbal mental age (VMA) estimate for each child. Syntactic skill was assessed using the Test for Reception of Grammar (TROG-2: Bishop, 2003) which requires picture-pointing responses to orally-presented sentences expressing a range of normatively emerging syntactic structures including relative clauses, pronoun agreement, and sentential embedding. With perfect success on four exemplars necessary to pass each structure, the test yields the raw scores (number of syntactic structures correct) that we used.

Pretests: Non-verbal mental ability. General (non-verbal) mental ability was measured using Raven's Progressive Matrices (Raven, 1989) a standardised, norm-referenced test requiring pointing to complete a graduated set of matrix patterns. Untimed standard (Raven, 1989) or coloured (Raven, 1962) versions were selected via the test manual. In the absence of relevant local norms, we used raw scores in all analyses.

Pre-/Post-tests: False belief. At pre-, post- and delayed-follow-up testing, children took four standard false belief trials at pre-test, post-test, and follow-up. None of these showed thought bubbles and stimuli were 3-dimensional unlike training stimuli. Baron-Cohen et al.'s (1985) Sally-Ann task was borrowed exactly with two trials and six questions (one test and two control per trial), performance on test and control questions was required to pass a trial so total scores ranged from 0 to 2. A two-trial misleading container false belief task assessed "near generalization." One trial involved a box containing a toy pig and the other box containing pencils. One test and two control questions had to be passed to pass each trial. Thus total false belief scores at each testing point (summing over trials) could range from 0 to 4.

Pre-/Post-tests: ToM Scale. At pre-, post- and delayed-follow-up testing, Wellman and Liu's (2004) five-step Guttman scale of sequenced ToM concepts (diverse desires, diverse beliefs, knowledge/ignorance, false belief and hidden emotion: see original publication for details) was administered and scored exactly as in Peterson, Wellman and Liu (2005) to yield a ToM Scale total score of 0 to 5. (We included the additional control question these authors used to guard against chance success on the hidden emotion task).

Timing of Pre-, Post- and Follow-Up Tests. The immediate post-test was administered at the end of completed training. All children mastered all six stages but rates of progress varied, similar to the original study. Depending on this, as well as days when children were available at school, multiple training sessions lasting 40 to 60 minutes each were spread either across several days (in 15 cases: 88%) or one long school day (in 2 cases: 12%). In each case, the ToM pretests were always given first. Next, after a break, children took the Raven's (1989) and language tests.

Then, after another longer break (including the lunch hour for the two with intensive one-day trainings), one or more multi-stage training sessions were administered, then the first post-test. All children managed to complete training in one, two or three sessions ($M = 2.12$, $SD = .70$) of 40 to 60 minutes each. The full set of training sessions encompassed a time period of between one and ten days ($M = 3.53$, $SD = 3.73$). On average, the post-test occurred a mean of 12.00 days ($SD = 11.22$) after pretesting. The delayed follow-up was a mean of 23.36 days ($SD = 23.16$) after the post-test (i.e. mean = 35 days after pretesting). (Note: Seven children in the training group, and one in the control group, were unavailable for delayed follow-up).

Results

Table 2 shows the mean scores for each group on pre-, post- and delayed-follow-up tests for: (a) the trained (changed location) false-belief task, (b) the total false belief battery (TFB), and (c) the 5-step ToM Scale. Within-group comparisons, conservatively using *Wilcoxon* matched-pairs tests, showed that, for the trained children, post-test scores were significantly higher than pretests for on all ToM variables. Specifically, for the trained task (a), $z = 3.27$, $p < .01$; for TFB (b), $z = 3.38$, $p < .01$; and for total scale steps (c), $z = 2.81$, $p < .01$. At the delayed follow-up, this significant improvement was consistently maintained, with the subgroup of trained children who took it scoring significantly higher than at pretest on all ToM measures: (a), $z = 2.43$, $p = .02$; (b), $z = 2.70$, $p < .01$; and (c), $z = 2.70$, $p < .01$. Control children, by contrast, performed no better at post-test than pretest on any ToM measure. Specifically, for (a), $z = -1.00$, $p = .32$, *ns*; for (b), $z = -1.73$, $p = .08$, *ns*; and for (c), $z = -1.93$, $p > .05$, *ns*. Likewise, at follow-up, the controls did not exceed pretest on any ToM variable: for (a), $z = -1.89$, $p > .05$, *ns*; for (b), $z = -1.67$, $p > .05$, *ns*; and for (c), $z = -1.13$, $p > .05$, *ns*.

Discussion

Replicating Wellman et al. (2002), findings trained children showed significant improvements in false belief test performance after thought bubble training. Extending upon this earlier study, this trained group outperformed a matched control group in pre- to post-test

comparisons (but not pretests). Furthermore, they generalized their improved understanding of minds to a developmental ToM Scale that included other ToM concepts (e.g., hiding true emotion) besides false belief. The durability of improvements for at least three weeks after training was also demonstrated for the first time. Thus it is now even clearer than before that Wellman et al.'s (2002) thought-bubble training scheme is an unusually effective one for boosting success on ToM tests by children with ASD.

Of course, it remains debatable whether these gains testify to a genuinely improved understanding of minds or merely “hacking out” of alternative ways to get false belief questions right (Happé, 1995). Our novel finding of robust gains on Wellman and Liu's (2004) ToM Scale after thought-bubble training adds further intrigue to this important theoretical debate. Whereas “hacking” heuristics can plausibly explain post-test gains by children with ASD on the false belief tasks they were trained on, it is harder to see how a “hacked-out solution ... which does not involve any appreciation of minds” (Happé, 1995, p. 852) could generalize far enough beyond them as to enhance performance on ToM-Scale tests like knowledge/ignorance or concealed emotion. Genuine gains in conceptual understanding of the mind's representational capacities might seem a more consistently plausible common alternative. However, it is important to emphasize that the gains we observed after thought bubble training were merely behavioral. Although post-test scores improved, we have no grounds to infer that our strategy taught children ToM understanding akin to that of their typically developing peers. Perhaps instead they gained a flexible alternative strategy for computationally solving a range of laboratory ToM tasks not limited to false belief. This is a worthy topic for further investigation. In this context, a crucial practical question for future research is whether gains made on ToM tests (achieved via development or via training) affect the social interactions of individuals with ASD, and if thought bubbles could serve a useful role in this extension beyond the laboratory.

While mechanism of learning behind the success of Wellman et al.'s thought bubble program is not yet clear, it would be intriguing to trial modified versions of it in applied social

settings. This might help to reveal whether or not real-life social skills and social cognition could benefit similarly from the kinds of stage-graded thought-bubble training these authors devised. In line with Fisher and Happé's (2005) call for further development of applied training programs to be systematically trialled by teachers and carers of individuals with autism, we believe our studies suggest clear future intervention potential in these areas. Both (a) the simple idea of using cartoon thought bubbles to represent thinking and (b) Wellman et al.'s detailed package of systematically sequenced training concepts ordered by increasing challenge, warrant being systematically evaluated. Although in its preliminary stages, we believe this new thought bubble procedure, more strongly validated now than previously, holds exceptional promise for being developed as a practical tool towards the ultimate goal of assisting individuals with autism to reach their full potential for social-cognitive development.

References

- Baron-Cohen, S., Leslie, A.M., & Frith, U. (1985). Does the autistic child have theory of mind? *Cognition*, 21, 37-46.
- Fisher, N., & Happé, F. (2005). A training study of theory of mind and executive function in children with autistic spectrum disorders. *Journal of Autism and Developmental Disorders*, 35(6), 757-771.
- Happé, F. G. E. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development*, 66(3), 843-855.
- McGregor, E., Whiten, A., & Blackburn, P. (1998). Teaching theory of mind by highlighting intention and illustrating thoughts: A comparison of their effectiveness with 3-year olds and autistic individuals. *British Journal of Developmental Psychology*, 16(3), 281-300.

- Peterson, C. C., Wellman, H. M., & Lui, D. (2005). Steps in theory-of-mind development for children with deafness or autism. *Child Development*, 76(2), 502-517.
- Swettenham, J., Baron-Cohen, S., Gomez, J. C., & Walsh, S. (1996). What's inside someone's head? *Cognitive Neuropsychiatry*, 1(1), 73 - 88.
- Wellman, H. M., Baron-Cohen, S., Caswell, R., Gomez, J. C., Swettenham, J., Toye, E., et al. (2002). Thought-bubbles help children with autism acquire an alternative to a theory of mind. *Autism*, 6(4), 343-363.
- Wellman, H.M., Cross, D. & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief: *Child Development*, 72, 655-684.
- Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, 75(2), 523-541.

Tables

Table 1. Means, (standard deviations) and statistical comparisons between groups on key variables at pre-test

Measure	Group Means		Group Difference	
	Control (<i>n</i> = 7)	Trained (<i>n</i> = 17)	<i>t</i> (22)	<i>p</i>
Age (months)	94.86 (28.69)*	79.41 (20.20)	1.51	.14
VMA (PPVT)	81.14 (33.99)	70.06 (21.31)	.97	.34
TROG-2 (raw) †	6.14 (5.46)	5.41 (2.81)	.44	.67
Raven's (raw)	6.14 (5.46)	5.41 (2.81)	.44	.65
Sally-Ann Total (out of 2)	.71 (.76)	.29 (.47)	1.66	.11
Total False Belief (out of 4)	1.86 (1.22)	1.18 (.73)	1.71	.10
ToM Scale Total (out of 5)	2.71 (1.11)	2.00 (.94)	1.61	.12

†As 7 children did not reach basal on TROG-2, raw scores replace VMAs

Table 2. Performance on pre-, post- and follow-up tests, by group

Group	Measure	Pre-Test Mean (SD)	Post-Test Mean (SD)	Follow-Up Mean (SD)
Experimental (<i>n</i> = 17)	Sally-Ann (out of 2)	.29 (.47)	1.53 (.80)	1.56 † (.73)
	Total False belief (out of 4)	1.18 (.73)	2.94 (1.25)	3.44† (.88)
	Total ToM Scale (out of 5)	2.00 (.94)	3.06 (1.00)	4.11† (.60)
Control (<i>n</i> = 7)	Sally-Ann (out of 2)	.71 (.76)	.57 (.79)	1.67†† (.82)
	Total False belief (out of 4)	1.86 (1.22)	1.43 (1.40)	3.00†† (1.55)
	Total ToM Scale (out of 5)	2.71 (1.11)	2.86 (1.68)	3.33†† (1.51)

† *n* = 9; †† *n* = 6