INTERNATIONAL JOURNAL OF COMPARATIVE PSYCHOLOGY

2018, 31
Charles Abramson
Special Issue Editor
Peer-reviewed

More than a Fluke: Lessons Learned from a Failure to Replicate the False Belief Task in Dolphins

Heather Hill [1], Sarah Dietrich [2], Alicia Cadena [1], Jenny Raymond [3], and Kyle Cheves [3]

[1] St. Mary's University, U.S.A.
[2] University at Buffalo, State University of New York, U.S.A.
[3] SeaWorld San Antonio, Inc., U.S.A

Critical to advanced social intelligence is the ability to take into consideration the thoughts and feelings of others, a skill referred to as Theory of Mind (ToM) or mindreading. In this article, we present a critical review of the comparative methodology and utility of the nonverbal FBT along with a description of an attempted FBT replication conducted with a bottlenose dolphin prior to the implementation of the more successful approaches used currently. Attempting to replicate Tschudin's (2001, 2006) methodology with dolphins highlighted several flaws that may explain the failures of socially complex mammals to display competency: (1) reliance on a containment invisible displacement procedure that is difficult for non-human animals and especially dolphins to follow, (2) a complex procedure that demands extensive training time, (3) a long trial duration with several moving parts which taxes the animal's memory and attention, and (4) a restricted number of two-choice FBT test trials, which limits statistical power given the small pool of trained animals. Although recent research paradigms for primates have corrected for some of these flaws, it is critical that comparative psychologists address these limitations for other species or taxa to be tested validly. Future research in ToM understanding through a false belief approach should move toward more ecologically valid designs and appropriate implicit measures that facilitate comparative approaches that can be replicated.

Dolphins have been popularly regarded as an especially clever species, with advanced intelligence and socioemotional lives (Fraser et al., 2006). Naturalistic observations of cooperative fishing, multi-level alliances, and long, stable relationships suggest that many species of dolphins live socially complex lives, which necessitate a high level of social intelligence (Connor & Krützen, 2015; Kuczaj, Tranel, Trone, & Hill, 2001). Critical to advanced social intelligence is the ability to take into consideration the thoughts and feelings of others, a skill referred to as Theory of Mind (ToM) or mindreading. In response to this special issue on comparative psychology today, the objective of this paper was to review the evidence and current comparative methods for measuring ToM with the nonverbal False Belief Task (FBT) designed by Call and Tomasello (1999) in primates and dolphins (Tschudin 2001, 2006).

Although recent studies appear to have corrected for the limitations of the original task designed by Call and Tomasello (1999) (i.e., Kano, Krupenye, Hirata, & Call, 2017; Krachun, Carpenter, Call, & Tomasello, 2009; Krupenye, Kano, Hirata, Call, & Tomasello, 2016), we present a critical review of the methodology and utility of the nonverbal FBT utilized by Tschudin (2001, 2006) to assess false belief understanding of bottlenose dolphins (*Tursiops truncatus*) in human care. Addressing the original methodological approach enables readers to interpret the results of a presented case study that was a direct replication attempt of Tschudin's original studies of the non-verbal FBT. This attempted replication elucidated a number of flaws that likely precluded primates and dolphins with advanced social cognition to pass FBTs and highlighted the importance of considering more ecologically valid testing paradigms, especially when a

Please send correspondence to Heather Hill, St. Mary's University, 1 Camino Santa Maria, San Antonio, Texas 78228. Email: hill1@stmarytx.edu https://doi.org/10.46867/ijcp.2018.31.01.13

comparative perspective is of interest; a topic tackled by several other papers in this special issue (e.g., Eaton et al., 2018; Smith, Watzek, & Brosnan, 2018; Zentall, 2018).

A Review of Comparative Research Investigating Theory of Mind

A developed ToM was defined originally by Premack and Woodruff (1978) as the point at which "an individual imputes mental states to himself or others (either to conspecifics or to other species as well)" (p. 515). Like many mental abilities, ToM is thought to be part of a larger "mindreading system" responsible for prerequisite skills, such as gaze following, joint attention, and declarative pointing (Baron-Cohen, 1995; Charman et al., 2000; Colonnesi, Rieffe, Koops, & Perucchini, 2008, see Table 1). According to the Machiavellian Intelligence Hypothesis, social competition encouraged the evolution of the intelligence necessary to develop complex social strategies (Byrne & Whiten, 1988). A similar hypothesis proposed by Dunbar and Schultz (2007), the social brain hypothesis, argued that the development of a complex brain along with more cognitively advanced skills (i.e., alliance formation, perspective taking, episodic memory, and ToM) was driven by the need to solve ecological conflicts through social cohesion. That is, individuals coordinate their own activities with others living in the same stable social group while satisfying both individual and group needs.

Table 1
Associated Cognitive Skills and Original Developmental Course for Theory of Mind in Humans as Measured by the False Belief Task in Research Conducted before 2000

Infancy	Birth - 18 mo	Social Perception	Joint Attention
Toddler-Early preschool	18 mo - 3 yr	Mental State Awareness Pass the True Belief Task (TBT)	Object Permanence (Piaget, 1951) Pretense & representation (Leslie, 1987) Situation theory (Perner, 1991)
Preschool	4 yr - 5 yr	Meta-representation Pass the False Belief Task (FBT)	Intentional deception Explanatory role of false beliefs (Wellman et al., 2001) Representation theory (Perner, 1991)
School Age	6 yr +	Recursion & Interpretation	

Note. This table represents the original work with the False Belief Task (FBT). There are a number of studies today that suggest that children less than 18 months may be able to pass modified non-verbal versions of FBT when using gaze duration and orientation (reviewed by Scott, 2017).

Comparative research on ToM has been largely primate-centric, focusing on identifying the evolutionary path toward uniquely human thought (see Table 2 for associated cognitive abilities across non-human species tested with the FBT). Bottlenose dolphins, while entirely separate from the primate evolutionary tree, share the evolutionary pressures of living in complex societies (Byrne & Whiten, 1988, or alternatively Dunbar & Schultz, 2007). This existence of complex societies and their subsequent complex social cognition is evidenced by the multi-level alliances exhibited by male dolphins (*Tursiops aduncus*) to hunt fish and herd female mates cooperatively (Connor, 2007; Connor & Krützen, 2003, 2015; Connor, Smolker, & Richards, 1992). Additionally, the shared knowledge of foraging strategies and its development appears to be passed between mothers and their offspring in some populations of bottlenose dolphins (*Tursiops aduncus*, Sargeant & Mann, 2009; Smolker, Richards, Connor, Mann, & Berggren, 1997) and Atlantic spotted dolphins (*Stenella frontalis*, Bender, Herzing, & Bjorkland, 2009). Social bonds between dolphins (spotted dolphins, bottlenose dolphins) are formed and mediated with pectoral flipper contact (Connor, Mann, & Watson-Capps, 2006;

Dudzinski, 1998; Dudzinski, Gregg, Paulos, & Kuczaj, 2010; Dudzinski, Gregg, Ribic, Kuczaj, 2009; Dudzinski & Ribic, 2017; Tamaki, Morisaka, & Taki, 2006). Finally, experimental research with bottlenose dolphins in managed care has demonstrated that the dolphins demonstrate long-term (up to 20 years) social recognition of conspecifics with whom they were once housed (Bruck, 2013).

Table 2
Associated Cognitive Skills and Original Developmental Course for Theory of Mind in Non-Human Animals as Measured by the False Belief Task

Cognitive Ability	Species Tested	Evidence	Citation
Attention	Chimpanzees	Observational & Experimental for all species	Reviewed by Leavens & Bard, 2011; Call & Tomasello, 2008
Joint Pointing	Orangutans		Reviewed by Leavens & Bard, 2011
	Dolphins		Xitco et al., 2004; Tomonaga & Uwano, 2010
Deception	Chimpanzees	Observational & Experimental for Apes	Byrne & Whiten, 1985; Hall & Brosnan, 2017; Whiten & Byrne, 1986; Suddendorf & Whiten, 2001
	Orangutans	Observational	Call & Tomasello, 1998; Suddendorf & Whiten, 2001
	Dolphins		Kuczaj et al., 2001; Miller, 2004
Empathy	Chimpanzees	Observational & Partial Experimental for Apes	Reviewed by Suddendorf & Whiten, 2001
	Orangutans	Observational	Reviewed by Suddendorf & Whiten, 2001
	Dolphins		Kuczaj et al., 2001
Visible Displacement	Chimpanzees	Experimental for all species	Call & Tomasello, 1999
	Orangutans		Call & Tomasello, 1999
	Dolphins		Jaakkola et al., 2010; Johnson et al., 2015, Singer & Henderson, 2015
Invisible Displacement	Chimpanzees	Observational & Experimental for all species	Reviewed by Suddendorf & Whiten, 2001
	Orangutans		Reviewed by Suddendorf & Whiten, 2001
	Dolphins		Johnson et al., 2015; Singer & Henderson, 2015
Mental Representations	Chimpanzees	Observational & Experimental for all species	Reviewed by Suddendorf & Whiten, 2001
	Orangutans		Reviewed by Suddendorf & Whiten, 2001
	Dolphins		Herman et al., 1999; Kuczaj et al., 2008; Pack, 2015; Pack & Herman, 2004
Social Intelligence	Chimpanzees	Observational & Experimental for all species	Call & Tomasello, 1998, 2008; Tomonaga et al., 2004
	Orangutans		Call & Tomasello, 1998, 2008
	Dolphins		Connor & Krutzen, 2003

In addition to facilitating cooperation, reasoning about others' thoughts can be used to deceive (Table 2). Dolphins in managed care have earned a reputation as "tricksters" primarily due to the number of anecdotes describing their attempts to manipulate trainers. In several independent accounts (described in Kuczaj et al., 2001), dolphins trained to retrieve objects foreign to their environment for a reward, selectively brought one item at a time, rather than the entire cache, presumably to maximize their rewards. In two other accounts, dolphins at separate facilities attempted to retrieve a conspecific's reward. After one dolphin performed a behavior, and before that dolphin returned to the trainer, an "imposter" dolphin appeared in front of the trainer. Adult dolphins appeared to perform this "deception" more often with novice trainers; one calf was observed

to attempt this same behavior in place of another similar-sized calf (Kuczaj et al., 2001). Although not empirical, these anecdotes and others like them have inspired curiosity into dolphin ToM (Table 2).

The ability to reason about others' thoughts develops in humans after the emergence of joint attention and imitation (Camaioni, 1992; Rogers & Pennington, 1991; Tomasello, 1995). Non-human animals follow similar developmental trends although pointing and gaze following appear to function slightly differently for dolphins (Table 1). Unlike primates, dolphins' eyes sit on either side of their head and can be used monocularly (using one eye to look at something) or binocularly (using both eyes to look at something). Dolphins tend to investigate stimuli monocularly but will position their bodies and heads to enable the use of both eyes (Blois-Heulin, Crével, Böye, & Lemasson, 2012).

Adapting methodology originating with human infants and extended to primates, trained and untrained dolphins can follow trainer's points, and in some studies, their gaze (Herman et al., 1999; Pack & Herman, 2004, 2007; Tschudin, Call, Dunbar, Harris, & van der Elst, 2001). Pack and Herman (2007) found that dolphins could follow a trainer's head turn but not the trainer's eye gaze and were most successful with larger rather than subtle head turns by the trainer. In a different study, dolphins attended to pointing but not to the trainer's head orientation (Tomonaga & Uwano, 2010). This variation in outcomes may have been due to different training reinforcement histories rather than cognitive variation; an issue that also occurs with primates (Eaton et al., 2018; Smith et al., 2018). In another study with a less controlled environment, two bottlenose dolphins spontaneously developed a pointing-like response in which the dolphins directed their trainers' attention to a particular object or location by positioning their body and rostrum in a specific direction (Xitco, Gory, & Kuczaj, 2001, 2004). The dolphins pointed more often when the trainers were looking than when their backs were turned, demonstrating a sensitivity to the attentional states of their trainers (Xitco et al., 2001, 2004).

A Measure of Theory of Mind - The False Belief Task

The measurement of cognitive abilities has been controversial for many years. While historically, the field of psychology was founded on measures of internal thoughts (i.e., introspective techniques used by Wundt and James), psychologists eventually moved to the perspective that behavior was the best representation to empirically measure psychological phenomena (reviewed briefly by Whissell, Abramson, & Barber, 2013). Psychology has once again experienced a shift in perspective with the study of cognitive processes once more at the forefront of psychology despite the difficulty in operationalizing the constructs consistently (see Whissell et al., 2013; Zentall, 2018 for brief discussions). A variety of cognitive processes have been studied extensively in humans and non-human animals (e.g., Smith et al., 2018; Wasserman, 2018; Zentall, 2018).

Ultimately, the majority of paradigms used to measure myriad cognitive skills in non-human animals is androcentric. For example, nonverbal tasks, measuring cognitive skills such as pointing, attention, or perspective-taking in human children have been adapted to test apes and chimpanzees from a comparative approach (Beran, 2017; Byrne & Whiten, 1988; Maestripieri, 2003). Only recently have these tasks been extended beyond primates (attention: Herman et al., 1999; Pack & Herman; 2004, 2007; Tschudin et al., 2001; object permanence: see review by Jaakkola, 2014; Johnson Sullivan, Buck, Trexel, & Scarpuzzi, 2015; Singer & Henderson, 2015; perspective-taking: see review by Pérez-Manrique & Gomila, 2017).

Original FBT with human children. The pivotal measure of ToM, the FBT, followed this path. First designed for children, it was adapted to non-human primates and was then extended to test bottlenose dolphins.

The task has been re-designed a number of times to compensate for a variety of issues but generally follows the following sequence of actions (i.e., Maxi task by Wimmer & Perner, 1983; Sally-Anne task by Baron-Cohen, Leslie, & Frith, 1985; and more recently, the location-change task, described by Call & Tomasello, 1999): a participant and an observer watch an apple placed on a table and then covered with a cup. When asked what the observer believed was under the cup, the participant could provide the correct answer merely by answering what he believed was under the cup. The participant's beliefs and the observer's beliefs would be overlapping. The FBT splits these beliefs apart by introducing a condition that creates a false belief in the observer by, for example, putting the apple in a cabinet after the observer walked out of the room. If asked the same question after the move, the participant would answer that the observer believed there was an apple under the cup, even though the participant knew it was in the cabinet. The FBT eliminates alternative explanations for success by decoupling what a participant knows is the state of the world from what a confederate is lead to believe falsely. Five-year old children consistently pass verbal FBTs while four-year old children struggle to achieve above chance accuracy (reviewed by Wellman, Cross, & Watson, 2001).

Non-verbal FBT with primates and human children. To compare false belief understanding between chimpanzees and orangutans directly, Call and Tomasello (1999) created a non-verbal adaptation of the traditional location change FBT. Both the verbal and nonverbal FBTs involved three roles in a hiding-finding task: an experimenter (sometimes called a hider), a communicator (sometimes called a confederate or an observer), and a participant.

The FBT begins when the experimenter hides a reward under one of two identical containers. The communicator can see where the reward is hidden, but the participant's view of the containers is blocked by a screen. The screen is removed and the communicator walks away. While the communicator is away, the experimenter switches the location of the containers. The communicator returns and indicates with a marker the container they "believe" holds the reward, then the participant is given a chance to choose a container. The participant is intended to work through the logic: (1) the communicator saw where the reward was placed and indicated that location; (2) the container was moved from where the communicator thought it was located; (3) the reward is therefore in the opposite location from where the communicator indicated. Descriptions of modifications to the primate version are detailed below and summarized in Table 3.

The task was designed to separate training general task requirements from the critical FBT probe trials. Previous designs had required that the primate participant receive feedback on repeated testing trials, leaving the studies vulnerable to associative explanations (Povinelli, Nelson, & Boysen, 1990; Woodruff & Premack, 1979). As Call and Tomasello (1999) wrote, "It could be argued that the reason apes took so long to learn in these tasks is that a certain number of trials are required to master general task demands and logistics concerning memory for the location of food, keeping track of people and objects simultaneously, and so forth. What is needed is a nonverbal task focused specifically on false belief understanding that gives participants some initial trials in which to master the general task requirements before the critical false belief task is given." (p. 382). Prior to the task, primates were provided with repeated exposure to habituate to trial elements (i.e., the apparatus, containers hiding food, the occlusion screen) and the role of the communicator. The final testing stage was divided into a pretest to demonstrate proficiency at using the communicator's marker to locate a reward, a short set of control trials, and then the critical FBT trials.

In the first set of training trials for the non-human primates, the subjects witnessed the communicator watching intently while items were moved behind a screen. Once the screen was removed, revealing two containers, the communicator lifted the correct container to momentarily reveal the food reward. The primate was then given the chance to select a container by touching it. All primate subjects passed the criterion of

above chance performance across two consecutive sessions after their first two sessions of 18 trials (total 36 trials per animal). The second set of training trials refreshed the primate's previous knowledge of markers to indicate containers holding rewards.

Table 3
Procedures Across Studies

	Call & Tomasello (1999) Study 1: Children	**Call & Tomasello (1999)** Study 2: Apes	**Tschudin et al. (1999, published in 2001)** Dolphins	**Tschudin (2006)** Dolphins	**Present Study** Dolphin
Training of Task					
Criteria	n/a	above chance within-session accuracy (10/18 trials) 2 consecutive sessions	not described	not described	proposed 80-85% achieved above chance within a session
Performance	n/a	2/2 orangutans & 2/2 chimps passed first 2 sessions (36 trials/animal)	not described	3/4 dolphins passed	No barrier: 115/183 (63%)
Marker Pretest					
Criteria	3 consecutive correct trials	above chance within-session accuracy (10/18 trials) 2 consecutive sessions	marker not used	marker not used	marker not used
Performance	28/28 4 & 5 yo after an average of 4-5 trials	2/2 orangutans & 2/2 chimps passed after 36 to 128 trials/animal	marker not used	marker not used	marker not used
Visible Displacement					
Criteria	50% accuracy on 2 trials	"above chance"	not conducted	not indicated	above chance
Performance	28/28	2/2 orangutans and 2/2 chimps passed	not conducted	satisfactory	No barrier: 3/5 (60%) Barrier + communicator stays: 2/2 (100%)
Visible Displacement (modified)					
Criteria	50% accuracy on 2 trials	"above chance"	not conducted	n/a	n/a
Performance	28/28	3/3 chimps passed	not conducted	n/a	n/a

/

Table 3 (continued)

	Call & Tomasello (1999)	Call & Tomasello (1999)	Tschudin et al. (1999/2001)	Tschudin (2006)	Present Study
Invisible Displacement					
Criteria	50% accuracy on 2 trials	"above chance"	unestablished	not indicated	above chance
Performance	10/14 4-year olds; 14/14 5-year olds	1/2 orangutans & 0/2 chimps passed	*Tt*1: 10/12 trials *Tt*2: 7/9 trials *Tt*3: 8/12 trials *Tt*4: 11/12 trials	satisfactory	no barrier: 3/5 (60%) barrier: 1/3 (33%)
Invisible Displacement (modified)					
Criteria	n/a	above chance accuracy on 4 trials (3/4)	n/a	n/a	n/a
Performance	n/a	1/1 orangutans & 5/5 chimps passed	n/a	n/a	n/a
Ignore Communicator					
Criteria	50% accuracy on 2 trials	not indicated	not indicated	not indicated	above chance
Performance	10/14 4-year olds; 14/14 5-year olds*	2/2 orangutans & 2/2 chimps passed	Jula: 10/12 trials Affrika: 8/8 trials Khanya: 10/12 trials Kani: 9/12 trials	not indicated	barrier: 1/3 (33%)
Ignore Communicator (modified)					
Criteria	n/a	above chance accuracy on 4 trials (3/4)	n/a	n/a	n/a
Performance	n/a	3/3 chimps passed	n/a	n/a	n/a
Nonverbal FBT					
Criteria	100% accuracy 4/4 trials within participant; > than chance success rates within each age category				
Performance	2/14 4 yo children & 8/14 5 yo children passed	0/2 orangutans & 0/2 chimps passed	Jula:7/8 trials Affrika: 4/4 trials Khanya: 4/4 trials Kani: 4/4 trials	Affrika: 5/6 trials Khanya: 2/3 trials Kani: 2/3 trials	barrier: 0/2 (0%)

Table 3 (continued)

	Call & Tomasello (1999)	Call & Tomasello (1999)	Tschudin et al. (1999/2001)	Tschudin (2006)	Present Study
True Belief Task					
Criteria	n/a	n/a	n/a	significantly > than chance across FBT and TBT per animal[b]	Modified (Dud-Tschudin, 2006), above chance
Performance	n/a	n/a	n/a	Affrika: 4/6 trials Khanya: 4/4 trials Kani: 2/3 trials	12/19 (63%)

Note. "above chance" was the only information provided in the articles; number of trials was not specified. [a]Call and Tomasello (1999) reported that of 14 4-year olds and 14 5-year olds, "During the control tests, three 4-year-old children failed both trials in the invisible displacement test and were dropped from the study. An additional 5- year-old child was also dropped because she became uncooperative during testing." (p. 385). However, FBTs are reported with 13 df for each age group. [b]True belief task had two forms: swap when communicator watched and no swap while communicator was away. Tschudin (2006) used both in the training phase, but it is not clear what was used in the testing phase.

Marker pretest trials followed the same procedure as the primate training trials, but rather than revealing the location of the reward, the communicator placed a marker on the correct container. Primates achieved the prerequisite criterion of above chance accuracy after two to five sessions of eight trials (total of 36 to 128 trials per animal). Four- and five-year-old children were expected to achieve a criterion of three consecutive correct trials and did so after experiencing four to five trials with most children doing so within that range.

Prior to exposure to the FBT, three controls were planned to test prerequisite knowledge: (1) visible displacement, (2) invisible displacement, and (3) ignore communicator (summarized in Table 3). These control trials differed slightly from a standard visible or invisible displacement task as the trials were set in a similar context to the experimental procedure: (1) the communicator marked the correct location, and (2) a time delay occurred between when the communicator left, after marking the correct location, and returned prior to the decision being made (see descriptions in Table 3). Most four-year old and all five-year old children passed the three controls (Call & Tomasello, 1999). All primates passed the visible displacement but only one primate met criteria for the invisible displacement control. For the three primates tested with the originally designed control tasks, the invisible displacement control, the ignore communicator control, and ultimately, the FBT were modified (i.e., the communicator placed a marker at the box/location of the reward and the marker stayed on the box during the swap. Once the tasks were modified, the remaining animals (n = 6) passed the controls (see descriptions in Table 3*)*.

Performance on the nonverbal-FBT for four-year old and five-year old children followed the established pattern found with verbal FBTs. Fewer four-year old children passed while a significant majority of five-year olds passed. None of the primates tested passed the FBT, even with the additional modification to assist with invisible displacement (see descriptions in Table 3). The results of this landmark study and introduction of the non-verbal FBT methodology were the standard by which subsequent studies were conducted, including the comparative replication and extension to non-primate animals such as dolphins (Tschudin, 2001, 2006).

Non-verbal FBT with dolphins. Bottlenose dolphins have a neocortex ratio that falls between the ratios of humans and non-human primates and correlates with increased degrees of sociality (i.e., social group size; Connor, 2007; Connor & Mann, 2006). These characteristics, along with the evidence for the presence of

other complex cognitive abilities found in non-human primates and human children (reviewed above, Tables 1 and 2), provided the justification for Tschudin (2001, 2006), to test dolphins on their capacity for social cognition, namely the ability to perceive the knowledge states of others using a non-verbal FBT. Despite the failures of great apes to pass the non-verbal FBT, Tschudin extended this methodology to bottlenose dolphins in managed care.

Tschudin's (2001) pilot study was a direct replication of Call and Tomasello's (1999) nonverbal FBT with two basic modifications. First, the dolphins observed a communicator tap the box containing the reward instead of leaving a marker for the dolphins to use as a cue. The original study by Call and Tomasello (1999) had suggested that the primates needed a visual cue to pass invisible displacement control trials. Interestingly, the dolphins in Tschudin's study successfully located the invisibly displaced reward on control trials without the assistive marker (Tschudin, 2001, 2006). Second, the dolphins indicated their response by orienting toward the location rather than having an opportunity to reach directly for the container. Tschudin (2001) recorded that four dolphins passed the non-verbal FBT, the first case of an animal passing the task.

Tschudin's seemingly landmark finding of false belief understanding in dolphins has been critiqued for ambiguity in reporting the results and the choice to publish in a book chapter rather than a peer-reviewed journal (see critique in Jaakkola, 2014). However, Tschudin's choice of publication outlet is not uncommon. Many of the initial attempts to capture false belief understanding in primates were also published primarily in book chapters (Premack, 1988; Hauser, 1999), as personal communications, or unpublished dissertations (O'Connell, 1996 as summarized by Povinelli & Giambrone, 2001).

The follow-up study (Tschudin, 2006), using the four original dolphins and a naïve, fifth dolphin, who after 166 training trials and inconsistent accuracy had testing terminated, added important controls, including a naïve experimenter and true belief trial. The third role, the naïve experimenter, was responsible for one task: the final release. After the trainer had baited the boxes, the communicator had indicated which box held the reward, and the trainer had performed the container manipulation (switching them or keeping them in their original locations), the naïve experimenter gave the "choose a side" signal. This modification was performed to control for any potential experimenter cues.

Tschudin (2006) interleaved two different types of true belief trials with false belief trials during testing. True belief trials were defined as trials where the communicator knew the true location of the reward (Table 3). In one true belief trial (TBT) variation, the communicator witnessed the switch. Just as in all trials, the baiter baited the box with a fish reward behind a screen as the communicator clearly watched, after which the communicator turned and walked away and then returned. Upon the communicator's return the two boxes were switched in front of both the dolphin and the communicator and then the communicator tapped which box held the fish when she left originally. In what Tschudin referred to as "dud" true belief control trials, the containers were not switched when the communicator was away (Tschudin et al., 2001). These two TBT controlled for unintentional communicator cues and determined if the dolphin differentiated between the communicator's awareness and unawareness of the container switch. Most importantly, the true belief swap trials decreased simple rule learning (i.e., if the containers are switched, choose the opposite side, Tschudin et al., 2001).

When the data for the four dolphins were pooled for the true belief trials, they succeeded on 10/13 (77% correct), which was significantly above chance (Table 3). Despite their successful performance on the TBTs with a naïve experimenter, the four dolphins failed to achieve a significant successful performance criterion at either the individual or group level (9/12, 75%).

A Case Study - Attempted Replication

In 2012, our research team initiated an effort to replicate Tschudin's follow-up experiments. The goal was to replicate as much of Tschudin's methodology as could be determined from the summary in his 2006 paper given that the training protocol used to teach the dolphins, the various aspects of the task, and the timeline it took to prepare the dolphins for the task, were not provided (Table 3). To fill any gaps in Tschudin's methodology, we followed the descriptions for the standardized testing procedure presented in Call and Tomasello (1999), including limiting the number of FBTs presented so as to control for the effect of experience. We wished to replicate the study while taking extra precautions to limit the possibility of associative learning through repeated exposure to the control or FBTs. As noted above, the step-by-step process of training tasks and individual animal performance is either excluded from methodology sections or minimized. We argue that excluding this information limits the replicability of the study and the possible interpretation of the final outcome (i.e., a behavior that emerged spontaneously or involved associative learning). The training methodology and outcomes for the attempted replication may be found in the following supplementary file.

In our attempt to replicate Tschudin (2001, 2006), two adult male bottlenose dolphins (*Tursiops truncatus*) and their trainers (5 and 20 years of experience, respectively) were selected. Both dolphins had basic training and were familiar with volunteering for husbandry behaviors (e.g., staying stationary in front of trainers; presenting body parts for inspection) and some performance behaviors (e.g., wave, jump into the air, spit water). Neither the dolphins nor the trainers were familiar with training complex cognitive concepts.

The trainers and research team created a plan for progressing through the stages of training and testing. An initial errorless discrimination process was planned to teach the dolphin to indicate the pole corresponding with the side (right or left) on which a fish was positioned. The dolphin was trained to touch the pole with his rostrum to indicate a choice and was given only one opportunity to make a choice (i.e., the clear response protocol). Following this step, other elements of the experimental tasks were trained with operant conditioning: the fish being placed into one of two opaque containers, the observer (i.e., communicator) indicating the correct side, a delay between the fish being presented and the decision, the observer coming and going from the training table, and the screen being moved to block and then reveal the containers. Special care was taken to create a plan that would expose the dolphin to the requirements of the task and the necessary experience to understand the task, while training the concept rather than simply training a behavioral response (Uyeyama, Pack, & Herman, 2005).

Following the philosophy behind the nonverbal FBT as outlined by Call and Tomasello (1999), training was isolated to the pre-testing stage so the animal was not given the opportunity for associatively learning the control or FBTs. All training and testing sessions were supervised by a primary researcher and videotaped. When training commenced, a criterion of 80-85% accuracy within a session was set as the prerequisite for testing trials, a traditional standard used in comparative cognitive research (e.g., Brown & Morrison, 1990; Morton, Lee, & Buchanan-Smith, 2013).

Training the first dolphin began April 2012 and did not progress as planned. First, session lengths ranged from five to 30 trials, depending on the attention of the dolphin and the training objectives for the day, making a percentage within trial accuracy less meaningful. For example, a session may have 2-3 errorless training trials (single pole on the side where the fish was positioned) followed by a training trial with both poles and the fish presented on the table and a final training trial with both poles and the fish presented in the

container. Accuracy criteria were adjusted to reflect the number of correct trials out of the total number of trials in that session (see Tschudin, 2001, 2006 for similar adjustments). Second, training the initial detection task, indicating the side with container holding the fish, took longer than expected (2 months of 327 trials across 17 sessions). Third, the dolphin's performance on trials fluctuated (Figure 1). While some variability is expected as the training progresses to more difficult steps, there were sessions in which the dolphin's performance was significantly below chance (below the thick red line in Figure 1). These unexpected session performances led to additional training trials in which we returned to earlier, easier steps. Due to these difficulties, training of the second dolphin was suspended and eventually halted so that resources (i.e., training session time) would not be split between the two animals. We summarize these training difficulties to highlight the possibility that Tschudin's original success may have been due to a "trained" behavior rather than a cognitive concept given the importance of training at every step of this task.

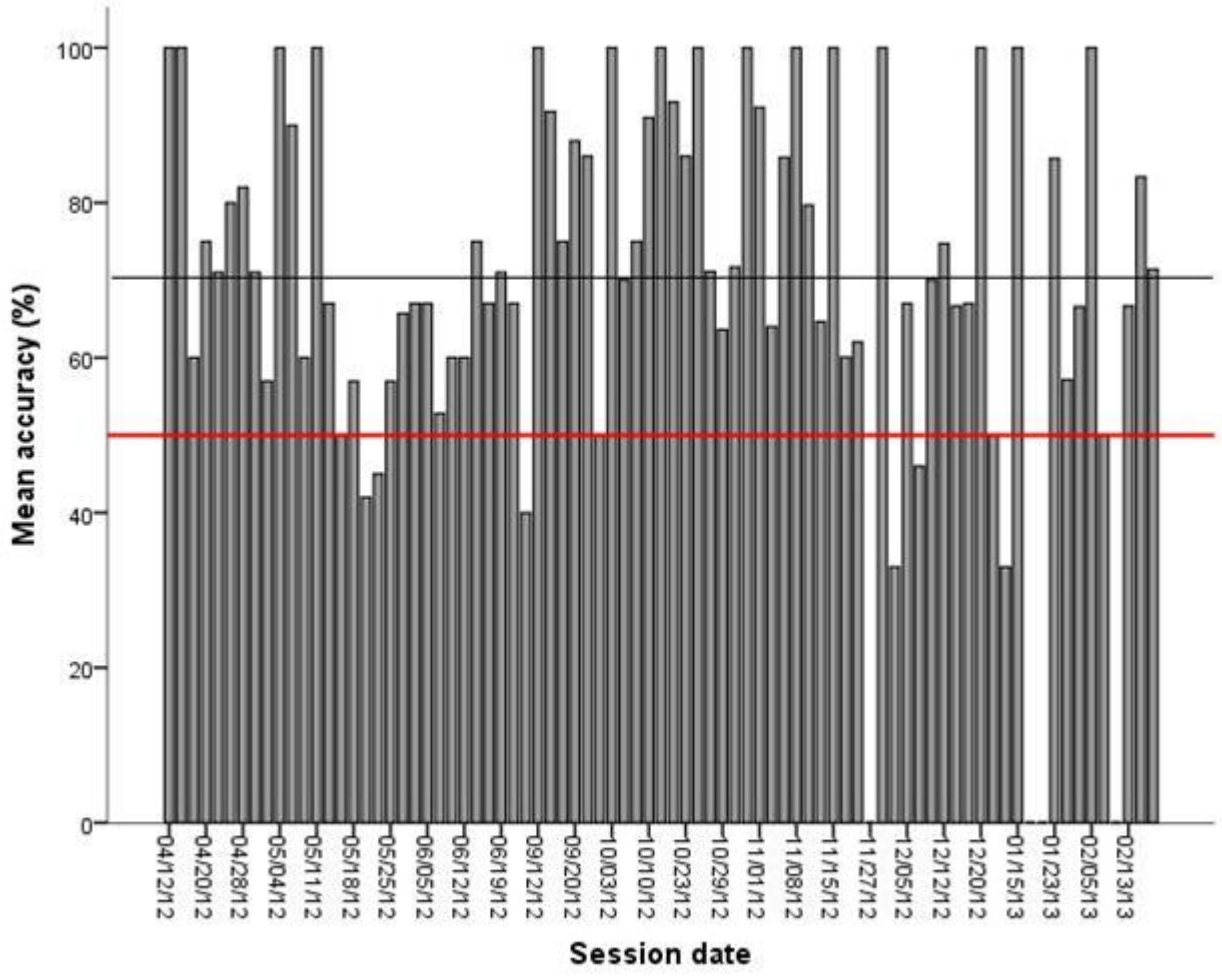

Figure 1. **Accuracy of responses across session**. The thick red line indicates chance performance. The black line indicates his overall level of accuracy. An examination of these below-chance sessions (range 0% - 45% correct) indicated that the sessions included trials in which the fish was fully visible (and should have been obvious) or involved trials that had been experienced previously and performed with accuracies ranging between 67% and 100%. These below chance performances had significantly more incorrect responses than would be expected by chance alone, $\chi^2(12, N = 622) = 41.95$, $p < 0.001$, Cramer's V = 0.26.

As we moved through the replication protocol, the importance of the assumptions that dolphins had object permanence and understood containers became evident. Partly to test these assumptions and partly to provide the dolphin experience, we ran the visible and invisible displacement trials. As performed by Tschudin (2001), we scheduled 12 trials of each type of displacement trial or testing. However, fewer trials were conducted in the final phases due to time and training constraints (Table 3). Ultimately, five visible displacement trials were conducted without a communicator, and the dolphin made a correct decision three times (60%). Of seven invisible displacement trials, the dolphin made a correct choice three times as well (43%), performing at chance.

Although our dolphin failed to achieve the pre-established accuracy levels on the preliminary trials, we decided to move forward with additional precursor FBT trials: visible and invisible displacement trials performed within the false belief context (e.g., the communicator was present and actively observing the baiting). These trials were longer than standard displacement trials, taxing both the animal's memory, attention span, and ability to track a reward's movement. Out of nine visible displacement trials with the communicator present, the dolphin chose correctly on six trials (67% accuracy). The dolphin made a correct decision on a single trial out of three tested for an invisible displacement with a communicator present.

Modified true belief trials ("dud trials" Tschudin, 2006) were performed with the dolphin making a correct choice on 12 out of 19 trials (63%). Finally, two FBTs were performed, although a total of four was planned per the previous protocols used (Call & Tomasello, 1999; Tschudin, 2001). The dolphin selected incorrectly on both trials despite making correct decisions on the last two invisible displacement trials that were conducted between the two false belief trials (see supplementary video). The final two FBTs were not conducted based on the failure rates and interest of the dolphin at the time of the sessions.

After a total of nine months with over 1000 training and probe trials, the dolphin never achieved the standard criterion set for acceptable performance (Table 3, Figure 1). Despite all of these trials, trying to train a concept- naïve dolphin resulted in a failure to demonstrate competency at invisible displacement of a fish in containers and to provide experimental evidence of ToM using the FBT. This experience in attempting to replicate the FBT highlighted the flaws in training, methodology, and reporting present in previous studies using the same methodology. These issues emphasize the common obstacles experienced when attempting replication and/or extension studies from a comparative perspective; tests must be designed so that when adapted to species with different characteristics the results are comparable (Eaton et al., 2018; Smith et al., 2018).

Discussion

Call and Tomasello (1999) created a nonverbal adaptation of the classic Sally-Anne location change FBT to directly compare the ability of children and non-human primates to reason about other's beliefs. To reduce simple rule learning, the training process necessary for animal participation was distinctly separated from control trials and the false belief test was restricted to four trials. Five-year old human children were able to successfully perform all stages of the task within a limited number of trials (i.e., an average of four pre-test experience trials, six control trials, and four FBT trials, Call & Tomasello, 1999) but great apes (i.e., orangutans and chimpanzees) were unsuccessful, never getting above 50% correct (Call & Tomasello, 1999). Tschudin (2001) reported near perfect performance by four bottlenose dolphins, but upon retesting with additional controls, the same four dolphins did not perform at levels significantly above chance when their data were pooled or tested individually (Tschudin, 2006).

In our attempted replication of Tschudin's study (2006), the bottlenose dolphin tested was unable to pass the invisible displacement trials, even after experience with visible displacement and what seemed an extensive training history. Despite sophisticated social structures and naturalistic displays suggesting advanced social cognition, attempts to capture ToM abilities of primates and dolphins using the original version of the nonverbal FBT (Call & Tomasello, 1999) have failed continuously. Overall, the difficulties encountered during a FBT replication attempt with a dolphin identified flaws in comparative design methodology, training, and reporting that produce questions about the validity of the task. The next section highlights several assumptions that underlie, and yet, may limit performance on the non-verbal FBT.

Pre-requisite Abilities Assumed: Invisible Displacement, Memory, Attention Span

Invisible displacement. The logic of Call and Tomasello's (1999) design hinges on the participant's ability to understand invisible displacement, specifically to track the location of a reward they had not seen hidden. As they wrote, "If subjects were not able to solve the problem of tracking the displacement of the food under these circumstances, they could not solve the false belief problem either" (p. 391). Although some great apes had to have a modified invisible displacement control in Call and Tomasello (1999), other apes have successfully passed switch invisible displacement trials with blinding as well as more complex tests of double invisible displacement (as reviewed by Jaakkola, 2014; Mallavarapu, Stoinski, Perdue, & Maple, 2014).

Tschudin's studies (2001, 2006) were some of the first tests of object permanence in dolphins. Prior to our replication attempt, Jaakkola, Guarino, Rodriguez, Erb, and Trone (2010) found that dolphins were unable to pass simple invisible displacement trials (e.g., the standard Piagetian or one where the containers cross paths but do not swap positions). Jaakkola et al. (2010) suggested that the dolphins' failures in invisible displacement tasks might be due to a constraint related to containment rather than an understanding of hidden movements, given the aquatic environment in which dolphins live and their use of echolocation to "see" through opaque surfaces.

Following these early invisible displacement attempts with dolphins, an alternative methodology was utilized by Johnson and her colleagues (2015). They coded behavioral responses of dolphins tracking an animated disk as it went "behind" occlusions. In a free swim scenario that required no training, the dolphins watched an animation reminiscent of a magician hiding an assistant. On the screen, an animated disk (the assistant) went behind a larger shape (the magician's cloak). The larger shape moved behind an occlusion, back into view, then burst in half, like a magician revealing that his cape is now empty. The dolphins looked toward the last occlusion that the larger shape had passed behind, suggesting that they understood that the magician's assistant (small disk) had been dropped off behind the screen. This modification is similar to many of the occlusion and violation of expectation tasks that have been successful (but see Heyes, 2014 who argues that these tasks with infants may be explained by a low-level novelty mechanism rather than an implicit ToM as critical controls have not been conducted) with infants as young as 13- (Surian, Caldi, & Sperber, 2007) and 15-months of age (Onishi & Baillergeon, 2005).

Overall, the invisible displacement of a hidden reward in containers is an essential aspect of the nonverbal FBT designed by Call and Tomasello (1999). Both great apes and dolphins appear to have object permanence depending on the testing technique. Primate species that succeeded at switch invisible displacement trials ultimately, failed to pass with the added complications of the FBT and required an assistive marker. For dolphins who appear to struggle with basic invisible displacement trials when containers are used, it is doubtful that they would pass the invisible displacement controls of the nonverbal FBT without relying on associative learning.

Memory. False belief task and control trials are relatively long (i.e., 2-3 minutes in length) and involve the coordination of several moving parts (the occluding screen and two containers) and people (the experimenter and communicator). Call and Tomasello (1999) adjusted the protocol they used with children to reduce the memory load on the non-human primates. The assistive markers were introduced with the rationale that the primates had trouble remembering the original location, preventing them from reasoning about the displacement. In the second adjustment, the researcher turned their back rather than leaving the room. As Call

and Tomasello (1999) wrote, “The setup of the experimental room did not allow the experimenter to leave the room quickly, potentially imposing an excessive memory load on the subjects” (p. 392).

In comparison, the design of the facilities allowed the communicator to leave the area and return for all dolphin replications (Tschudin 2001, 2006 and the presented replication). Previous tests of the nonverbal FBT did not report the duration of their control and testing trials. Trials in the replication took about 20-30 seconds from the initial reward placement until the release. This duration exceeds the duration adult humans can actively recall information from their working memory without opportunities for rehearsal (Brown, 1958; Peterson & Peterson, 1959). If four- and five-year-olds do not spontaneously use rehearsal to remember information, it is highly questionable that primates or dolphins will spontaneously rehearse the steps they just observed in the FBT much less remember the relevant details from start to finish over the duration of the trial. On visual match-to-sample trials, dolphins have maintained near perfect performance up to 30 seconds with accuracy gradually dropping but remaining around 70% for up to 80-second delays. Even the best dolphin's ability to remember and mimic a trainer's actions was shorter than for visual object matching, around 85% for 30-second delays and dropped off at 80-second delays to 60% accuracy (Herman, 2002).

It is plausible that the FBT control and testing trials may have taxed the memory of both the primates and dolphins. Comparatively, both can pass short-term memory tasks with relatively high accuracy in similar task durations. The FBT procedure, unlike memory research, is also layered with several complex cognitive skills. The impact of this layered complexity on dolphin working memory is uncertain, making it difficult to determine to what extent working memory, object permanence, and false belief understanding were tested in this paradigm. Simply put, if dolphins cannot remember the task from start to finish, they cannot follow the logic of the task.

Ecological validity. The traditional Sally-Anne verbal FBT is clearly an inappropriate tool for measuring false belief in non-human animals. Some tasks are also ecologically inappropriate, as illustrated humorously by an article in the *Onion*, “Study: Dolphins Not So Intelligent On Land.” Unfortunately, determinations of ecological appropriateness for less obviously inappropriate tasks require knowledge of a species’ sensory abilities, social groupings, environmental demands, and typical behavioral repertoires.

After failing the FBT designed by Call and Tomasello (1999) repeatedly, chimpanzees were much more successful once the task was re-designed to capitalize on the chimpanzees’ natural competitive social tendencies. Chimpanzees were more successful when making decisions based on the awareness of either a dominant or subordinate competitor’s knowledge but maybe not a full understanding of false belief (Hare, 2001; Hare, Call, & Tomasello, 2001; Kaminski, Call, & Tomasello, 2008). Like the studies performed with human infants (review by Baillargeon Scott, & He, 2010 and critique by Heyes, 2014), once implicit measures of understanding, namely gaze direction (Krachun et al., 2009) and anticipatory gaze measures (Kano et al., 2017; Krupenye et al., 2016, but see Andrews, 2017) were utilized, chimpanzees were more likely to pass the FBT. Clearly, alternative methodologies may facilitate a comparative perspective.

The sensory abilities, environmental demands, and typical behavioral repertoires of dolphins are distinct from primates. Some basic modifications were indisputably necessary to begin replicating the primate nonverbal FBT. Where primates grabbed for the location of the reward, dolphins had to either orient at the location (original study summarized in Tschudin, 2001) or tap a corresponding target (Study 2, Tschudin, 2001, 2006, our replication). Other non-modifications of the procedures introduced variation. For example, the primates and researchers were across from each other but separated by the containment wall. The testing

apparatus for dolphins was similarly placed at the end of the enclosure, leaving dolphins with a different experience of looking up from aquatic environment at the researchers.

In summary, the nonverbal FBT is inherently primate-centric, much like the original FBT was androcentric. Some latent assumptions (e.g., cooperation and containment) have been isolated and ecologically appropriate adjustments have yielded more promising evidence for FBT understanding in chimpanzees. We suggest that to facilitate the possibility for comparative designs to demonstrate false belief understanding by dolphins and other animals, more ecologically appropriate methods and measures are needed as the originally designed location change, nonverbal FBT is not an effective method. Thus, any replication attempts would most likely be unsuccessful.

Training concerns. In any experiment in which a trained behavior is required for a response, threats to internal validity are imbedded in the design as external cues can be given or trained inadvertently at any time during the training. For many comparative studies, training is time intensive and vulnerable to cueing if double-blind studies are not conducted or over-training occurs. In most studies, the training history is rarely explained or videotaped and presented as part of the final results. While many methodologies use established protocols that are written for replication purposes, less established methodologies are subject to interpretation if adequate descriptions of training procedures are not provided.

Our attempted replication was subject to this ambiguous methodology issue, particularly with regard to the training of the basic protocol. Very limited information was provided by Tschudin's (2001, 2006) descriptions of FBT training, which produced a number of inferences about the necessary training steps. Without knowing the training utilized, Tschudin's pilot study (2001) results could have been due to a Clever Hans effect, and the results of his subsequent study (2006) may have been related to repeated experience with the task. The critical flaw in Tschudin's original protocol for conducting the study was the lack of a double-blind procedure (i.e., the experimenter both baited and presented the containers for the dolphins to choose between once the communicator had indicated the container). This issue was addressed in the follow-up study by introducing a naïve assistant to present the containers to the dolphin for a choice. However, since the training history for each dolphin was not reported or apparently recorded and probe trials were not conducted, it is very possible that the trainers trained the task rather than the concept. Training a task, such as learning to deceive a human participant only after extensive training, does not support the goal of demonstrating a cognitive ability based on its natural acquisition. Rather, this extensive training introduces discriminative stimuli that can be used to pass a task without cognitive understanding (Gallup, Anderson, & Schillito, 2002; Heyes, 1993; Povinelli & Giambrone, 2001).

Design Issues of Non-verbal False Belief Task (Call & Tomasello, 1999)

When to hold or when to fold? One of the major issues facing science today is the importance of replication, especially of cognitive and social psychological concepts (e.g., Maxwell, Lau, & Howard, 2015). In fact, as a basic tenet of the scientific method, every effort should be made to replicate previous research. Unfortunately, being second to the finish line is often not as valued as contributing to the scientific literature. However, in some cases replication may be virtually impossible, or at the very least, very difficult due to limited subjects, access, training time, or resources as is the case for much of comparative psychology. These issues are particularly pertinent for dolphin cognitive research, which currently finds itself in a replication bias

or crisis state. First, many of the original cognitive studies were conducted with a handful of dolphins in a limited number of laboratories (Herman, 2010; Pack, 2015). Since these original studies, diversity across the facilities that conduct cognitive research in odontocetes has increased but is still very limited to a handful of researchers and limited subjects. Concentrating studies with a small number of animals has several advantages, including "savvy" animals, or animals that have learned how to learn after 1000s of trials over time, knowledgeable trainers, and dedicated training time. Unfortunately, this limited resource also means that replication studies are more difficult given the trouble in publishing successful replication studies, much less a replication study that is counter to an accepted finding, or a replication attempt that failed during training of the prerequisite skills because of limited power or invalid methodological tasks.

Power issues. One of the largest flaws in replication designs is sufficient power (Maxwell et al., 2015), which is one reason meta-analyses are so helpful (e.g., Wellman et al., 2001). The FBT is fraught with power issues because of the threat of experience to interpretation of performance on FBT trials. Most FBT studies perform four to eight trials of the actual FBT to minimize the possibility of associative learning. Unfortunately, this minimum number of trials does not leave much room for errors at an individual level and requires large sample sizes if significance is to be obtained. In Call and Tomasello's (1999) initial study, the primates were moved forward in their training if they correctly selected on a majority of the trials in two consecutive sessions. This ability to make mistakes or not achieve a standard criterion (i.e., 80-85% accuracy that is often used in comparative cognitive studies) was then expected to culminate into near perfect performance on a more complex task with fewer trials, a seemingly impossible task. While seven subjects seems like a lot of primates, using two choices and four trials each on the FBT is not reasonable to achieve statistical power. If each animal missed one trial, individual performances would never reach statistical significance despite doing so if data are pooled.

With this power issue in mind, an examination of Tschudin's study factors is necessary. Based on the number of dolphins that were deemed plausible candidates (N = 5, 1 excluded for poor performance), the number of test trials used (1999: N = 4, 2001: N > 4), which was minimal to limit role of experience/learning in the task, highly repetitious training trials so FBT trials can be performed with a limited number, and the number of options to choose between (chance = 50%), there was not enough power to achieve statistical significance at either the individual level or group level unless every dolphin performed with near 100% accuracy. Whether primates or dolphins, is it reasonable to ask a non-human to perform perfectly on a cognitive task with only four attempts to demonstrate the concept while simultaneously hoping that a successful first trial performance can be used as evidence for a novel transfer task? Moreover, is it reasonable to draw this conclusion if the elicitation of this spontaneous abstract ability occurred after intense training on aspects of the paradigm, which are needed to conduct the final test?

Conclusion

Despite persistent investigations into ToM understanding using Call and Tomasello's (1999) nonverbal FBT, primates have never passed the task and dolphins who passed in one study (Tschudin, 2001), failed after additional controls were introduced (Tschudin, 2006). Our experience attempting to replicate Tschudin's (2001, 2006) tests with dolphins highlighted several flaws in methodology that may explain why socially complex mammals fail to display competency: (1) reliance on a containment invisible displacement procedure that is difficult for non-human animals, and especially dolphins, to follow, (2) a complex procedure which demands extensive training, (3) a long trial duration with several moving parts, taxing an animal's

memory and attention, and (4) a restricted number of two-choice FBT test trials, which limits statistical power given the small pool of trained animals. After reviewing Tschudin's methodology and results critically, statistical power most likely prevented the results in the second study from reaching statistical significance, but we cannot rule out the possibility the dolphins learned the task associatively and were unable to pass due to the additional true belief control trials. While adding additional choices (i.e., more than two) would increase the statistical power and facilitate statistical analyses without increasing the trials or number of animals that need to be trained and tested, it would also complicate the task further by requiring that the subject keep track of more than two locations.

Comparative research with primates has demonstrated that naturalistic designs using species-appropriate tasks that were meaningful to the animal (e.g., chimpanzee competition design) were effective. Additionally, creating methods that elicit spontaneous, implicit measures of understanding such as anticipatory gazes or frequency of gazes appear promising for all species currently tested (e.g., human infants, chimpanzees). However, as has been cautioned by a number of researchers, low-level associative explanations must be controlled for in these studies (e.g., Andrews, 2017; Heyes, 2014). Finally, experimental procedures that harness ecological validity have been more successful in demonstrating advanced aspects of ToM and complex social behaviors than the more artificial procedures in primates and dolphins (e.g., Buttelmann, Buttelmann, Carpenter, Call, & Tomasello, 2017; Hare, 2001; Johnson et al., 2015; Kaminski et al., 2008; Kano et al., 2017; Krupenye et al., 2016). For example, Johnson et al. (2015) successfully modified the traditional container testing paradigm for invisible displacement with dolphins by adapting an occlusion method and gaze direction measure successfully utilized with human infants, which evidences the need for an adaptable comparative approach (Onishi & Baillergeon, 2005; Surian, Caldi, & Sperber, 2007). The training and testing protocols used by Brosnan and her colleagues in their investigations of cooperation, decision making, and equity using variations of the "Stag Hunt" game or the Assurance game are excellent examples of ecologically valid methodologies that require limited training and enable cross-species comparisons (Brosnan et al., 2011).

When we designed our FBT study, our goal was to replicate the study conducted by Tschudin (2001, 2006) using the same methodology. For this reason, we did not seek to modify or improve the methodology. As we have ultimately discovered, the FBT task originally designed by Call and Tomasello (1999) was not ecologically valid for the primates and instead tested memory, attention, and object permanence at the highest Piagetian level (i.e., invisible displacement), critical flaws identified repeatedly by scientists who argued that this approach was overly complex and relied on additional cognitive skills needed to track all the elements involved (Bloom & German, 2000; Wellman et al., 2001). In the last several years, naturalistic designs with implicit measures of performance have produced stronger empirical support for primates passing false belief tests (Buttelmann et al., 2017; Kaminski et al., 2008; Kano et al., 2017; Krupenye et al., 2016, but see Andrews, 2017).

We issue a call for action that similar modifications be made for dolphins, as failing the current version of the FBT does not preclude the existence of understanding of other's attributions nor of ToM by dolphins. To assess false belief attributions in dolphins, future study designs should consider relevant life history and sensory characteristics and eliminate the need for complex training and testing using artificial procedures, echoing Delfour's message from 2010. A FBT paradigm might utilize a cooperation-competition context in which animals acquire information about an object or a set of objects through echolocation and eavesdropping on one another that sets up a series of contingencies that may be manipulated or modified such that one animal benefits but not the other or both benefit when a choice is made. If this experimental set-up can be automated, then the influence of humans is removed and any outcome is less confounded. Increasingly, research designs

are becoming more innovative as the perspective and abilities of the various subjects are recognized and incorporated (i.e., Abramson, Dinges, & Wells, 2009; Buttelmann et al., 2017; Johnson et al., 2015) and will be more successful and valid in demonstrating whether or not specific cognitive abilities exist.

Acknowledgments

We would like to sincerely thank SeaWorld San Antonio and the scientific review committee for providing the trainers, training time, and dolphins to conduct this study. We especially appreciate the continued support of Chris Bellows and Dr. Judy St. Leger in our research endeavors to learn what we can about how marine mammals view their world. Many thanks to Sara Guarino for her editorial assistance. Finally, we appreciate the suggestions provided by Dr. David Washburn and several anonymous reviewers on earlier drafts.

References

Abramson, C., Dinges, C., & Wells, H. (2009). Operant conditioning in honey bees (*Apis mellifera* L.): The cap pushing response. *PLOS One, 11*, e0162347. doi:10.1371/journal.pone.0162347.

Andrews, K. (2017). Apes track false beliefs but might not understand them. *Learning & Behavior*, 1–2. https://doi.org/10.3758/s13420-017-0288-8

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14*, 110–118.

Baron-Cohen, S. (1995). *Mindblindedness.* Boston, MA: MIT Press.

Baron-Cohen, S., Leslie, A., & Frith, U. (1985). Does the autistic child have a "Theory of mind"? *Cognition, 21*, 37–46.

Bender, C. E., Herzing, D. L., & Bjorklund, D. F. (2009). Evidence of teaching in Atlantic spotted dolphins (*Stenella frontalis*) by mother dolphins foraging in the presence of their calves. *Animal Cognition, 12*, 43–53.

Beran, M. J. (2017). To err is (not only) human: Fallibility as a window into Primate cognition. *Comparative Cognition & Behavior Reviews, 12*, 57–81.

Blois-Heulin, C., Crével, M., Böye, M., & Lemasson, A. (2012). Visual laterality in dolphins: Importance of the familiarity of stimuli. *BMC Neuroscience, 13*, 1–9.

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as test of theory of mind. *Cognition, 77*, B25–B31.

Brosnan, S. F. (2018). The importance of a truly comparative methodology for comparative psychology. *International Journal of Comparative Psychology, 31*, 1–5.

Brosnan, S. F., Parrish, A., Beran, M. J., Flemming, T., Heimbauer, L., Talbot, C. F., ... Wilson, B. J. (2011). Responses to the Assurance game in monkeys, apes, and humans using equivalent procedures. *Proceedings of the National Academy of Sciences, 108*, 3442–3447.

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology, 10*, 12-21.

Brown, M. F., & Morrison, S. K. (1990). Element and compound matching-to-sample performance in pigeons: The roles of information load and training history. *Journal of Experimental Psychology: Animal Behavior Processes, 16*, 185–192.

Bruck, J. N. (2013). Decades-long social memory in bottlenose dolphins. *Proceedings of the Royal Society of London B: Biological Sciences, 280*, 20131726.

Buttelmann, D., Buttelmann, F., Carpenter, M., Call, J., & Tomasello, M. (2017). Great apes distinguish true from false beliefs in an interactive helping task. *PLoS One, 12*, e0173793.

Byrne, R. W., & Whiten, A. (1988). Towards the next generation of data quality – A new survey of primate tactical deception. *Behavioral and Brain Sciences, 11*, 267–271.

Call, J., & Tomasello, M. (1999). A nonverbal false belief task: The performance of children and great apes. *Child Development, 70*, 381–395.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences, 12*, 187–192.

Camaioni, L. (1992). Mind knowledge in infancy: The emergence of intentional communication. *Infant and Child Development, 1*, 15–22.

Charman, T., Baron-Cohen, S., Swettenham, J., Baird, G., Cox, A., & Drew, A. (2000). Testing joint attention, imitation, and play as infancy precursors to language and theory of mind. *Cognitive Development, 15*, 481–498.

Colonnesi, C., Rieffe, C., Koops, W., & Perucchini, P. (2008). Precursors of a theory of mind: A longitudinal study. *British Journal of Developmental Psychology, 26*, 561–577.

Connor, R. C. (2007). Dolphin social intelligence: Complex alliance relationships in bottlenose dolphins and a consideration of selective environments for extreme brain size evolution in mammals. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *362*, 587–602.

Connor, R. C., & Krützen, M. (2003). Levels and patterns in dolphin alliance formation. In F. deWaal & P. Tyack (Eds.), *Animal social complexity: Intelligence, culture and individualized societies* (pp. 115–120). Boston, MA: Harvard University Press.

Connor, R. C., & Krützen, M. (2015). Male dolphin alliances in Shark Bay: Changing perspectives in a 30-year study. *Animal Behaviour*, *103*, 223–235.

Connor, R. C., & Mann, J. (2006). Social cognition in the wild: Machiavellian dolphins?. In S. Hurley & M. Nudds (Eds.), *Rational animals?* (pp. 329–367). Oxford, UK: Oxford University Press.

Connor, R., Mann, J. & Watson-Capps, J. (2006). A sex-specific affiliative contact behavior in Indian Ocean bottlenose dolphins, *Tursiops* sp. *Ethology, 112,* 631–638.

Connor, R. C., Smolker, R. A., & Richards, A. F. (1992). Dolphin alliances and coalitions. In A. H. Harcourt & F. B. M. de Waal (Eds.), *Coalitions and alliances in humans and other animals*, (pp. 415–443). Oxford, UK: Oxford University Press.

Delfour, F. (2010). Marine mammals enact individual worlds. *International Journal of Comparative Psychology*, *23*, 792–810.

Dudzinski, K. M. (1998). Contact behavior and signal exchange in Atlantic spotted dolphins (*Stenella frontalis*). *Aquatic Mammals, 24*, 129–142.

Dudzinski, K., Gregg, J., Paulos, R. D., & Kuczaj, S. A. II (2010). A comparison of pectoral fin contact for three distinct dolphin populations. *Behavioural Processes, 84,* 559–567.

Dudzinski, K., Gregg, J., Ribic, C., & Kuczaj, S. (2009). A comparison of pectoral fin contact between two different wild dolphin populations. *Behavioural Processes, 80,* 182–190.

Dudzinski, K. M., & Ribic, C. A. (2017). Pectoral fin contact as a mechanism for social bonding among dolphins. *Animal Behavior and Cognition, 4*, 30–48.

Dunbar, R. I., & Shultz, S. (2007). Evolution in the social brain. *Science*, *317*, 1344–1347.

Eaton, T., Hutton, R., Leete, J., Lieb, J., Robeson, A., & Vonk, J. (2018). Bottoms-up! Rejecting top-down human-centered approaches in comparative psychology. *International Journal of Comparative Psychology, 31*, 1–6.

Fraser, J., Reiss, D., Boyle, P., Lemcke, K., Sickler, J., Elliott, E., ... Gruber, S. (2006). Dolphins in popular literature and media. *Society & Animals, 14*, 321-349.

Gallup, G. G., Anderson, J. R., & Shillito, D. J. (2002). The mirror test. In M. Bekoff, C. Allen, & G. Burghardt (Eds.), *The cognitive animal* (pp. 325–333). Cambridge, MA: MIT Press.

Hall, K., & Brosnan, S. F. (2017). Cooperation and deception in primates. *Infant Behavior and Development, 48*, 38-44.

Hare, B. (2001). Can competitive paradigms increase the validity of experiments on primate social cognition? *Animal Cognition, 4*, 269–280.

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour*, *61*, 139–151.

Hauser, M. (1999). Primate representations and expectations: Mental tools for navigating in a social world. In P. Zelazo, J. Astington, & D. Olson (Eds.), *Developing theories of intention: Social understanding and self-control* (pp. 169–194). Mahwah, NJ: Lawrence Erlbaum.

Herman, L. M. (2002). Exploring the cognitive world of the bottlenosed dolphin. In S. J. Armstrong & R. G. Botzler (Eds.), *The animal ethics reader*, 1st ed. (pp. 161–165). London, UK: Routledge.

Herman, L. M. (2010). What laboratory research has told us about dolphin cognition. *International Journal of Comparative Psychology, 23*, 310–330.

Herman, L. M., Abichandani, S. L., Elhajj, A. N., Herman, E. Y., Sanchez, J. L., & Pack, A. A. (1999). Dolphins (*Tursiops truncatus*) comprehend the referential character of the human pointing gesture. *Journal of Comparative Psychology, 113*, 347–374.

Heyes, C. M. (1993). Anecdotes, training, trapping and triangulating: do animals attribute mental states?. *Animal Behaviour*, *46*, 177–188.

Heyes, C. (2014). False belief in infancy: A fresh look. *Developmental Science, 17*, 647–659.

Jaakkola, K. (2014). Do animals understand invisible displacement? A critical review. *Journal of Comparative Psychology, 128*, 225–239.

Jaakkola, K., Guarino, E., Rodriguez, M., Erb, L., & Trone, M. (2010). What do dolphins (*Tursiops truncatus*) understand about hidden objects? *Animal Cognition*, *13*, 103–120.

Johnson, C. M., Sullivan, J., Buck, C. L., Trexel, J., & Scarpuzzi, M. (2015). Visible and invisible displacement with dynamic visual occlusion in bottlenose dolphins (*Tursiops* spp). *Animal Cognition, 18*, 179–193.

Kaminski, J., Call, J., & Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition*, *109*, 224–234.

Kano, F., Krupenye, C., Hirata, S., & Call, J. (2017). Eye tracking uncovered great apes' ability to anticipate that other individuals will act according to false beliefs. *Communicative & Integrative Biology, 10*, 1–12.

Krachun, C., Carpenter, M., Call, J., & Tomasello, M. (2009). A competitive nonverbal false belief task for children and apes. *Developmental Science, 12*, 521–535.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*, 110–114.

Kuczaj, S., Tranel, K., Trone, M., & Hill, H. (2001). Are animals capable of deception or empathy? Implications for animal consciousness and animal welfare. *Animal Welfare, 10*, 161–173.

Leavens, D. A., & Bard, K. A. (2011). Environmental influences on joint attention in great apes: Implications for human cognition. *Journal of Cognitive Education and Psychology*, *10*, 1-23.

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind". *Psychological Review, 94*, 412–426.

Maestripieri, D. (2003). The past, present, and future of primate psychology. In D. Maestripieri (Ed.), *Primate psychology* (pp. 1–16). Cambridge, MA: Harvard University Press.

Mallavarapu, S., Stoinski, T. S., Perdue, B. M., & Maple, T. L. (2014). Double invisible displacement understanding in orangutans: Testing in non-locomotor and locomotor space. *Primates*, *55*, 549–557.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean?. *American Psychologist, 70*, 487.

Miller, A. (2004). *Social manipulation in the bottlenose dolphin: A study of deception and inhibition.* Retrieved from Scholar Space. http://hdl.handle.net/10125/11868

Morton, F. B., Lee, P. C., & Buchanan-Smith, H. M. (2013). Taking personality selection bias seriously in animal cognition research: A case study in capuchin monkeys (*Sapajus apella*). *Animal Cognition, 16*, 677–684.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*, 255–258.

Pack, A. A. (2015). Experimental studies of dolphin cognitive abilities. In D. L. Herzing & C. M. Johnson (Eds.), *Dolphin communication and cognition: Past, present, and future* (pp. 175–200). Cambridge, MA: MIT Press.

Pack, A. A., & Herman, L. M. (2004). Bottlenosed dolphins (*Tursiops truncatus*) comprehend the referent of static and dynamic human gazing and pointing in an object-choice task. *Journal of Comparative Psychology, 118*, 160–171.

Pack, A. A., & Herman, L. M. (2007). The dolphin's (*Tursiops truncatus*) understanding of human gazing and pointing: Knowing what and where. *Journal of Comparative Psychology*, *121*, 34–45.

Pérez-Manrique, A., & Gomila, A. (2017). The comparative study of empathy: sympathetic concern and empathic perspective-taking in non-human animals. *Biological Reviews, 93*, 248–269.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: The MIT Press.

Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*, 193–198.

Piaget, J. (1951). *The child's conception of the world* (No. 213). Lanham, MD: Rowman & Littlefield.

Povinelli, D. J., & Giambrone, S. (2001). Reasoning about beliefs: A human specialization? *Child Development*, *72*, 691–695.

Povinelli, D. J., Nelson, K. E., & Boysen, S. T. (1990). Inferences about guessing and knowing by chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology*, *104*, 203–210.

Premack, D. (1988). `Does the chimpanzee have a theory of mind?' revisited. In R. Byrne & A. Whiten (Eds.), *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans* (pp. 160–179). Oxford, UK: Oxford University Press.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences, 1*, 515–526.

Rogers, S. J., & Pennington, B. F. (1991). A theoretical approach to the deficits in infantile autism. *Development and Psychopathology, 3*, 137–162.

Sargeant, B. L., & Mann, J. (2009). Developmental evidence for foraging traditions in wild bottlenose dolphins. *Animal Behaviour, 78*, 715–721.

Scott, R. M. (2017). The developmental origins of false-belief understanding. *Current Directions in Psychological Science, 26*, 68–74.

Singer, R., & Henderson, E. (2015). Object permanence in marine mammals using the violation of expectation procedure. *Behavioural Processes, 112*, 108–113.

Smith, M. F., Watzek, J., & Brosnan, S. F. (2018). The importance of a truly comparative methodology for comparative psychology. *International Journal of Comparative Psychology, 31*, 1–16.

Smolker, R. A., Richards, A. F., Connor, R. C., Mann, J., & Berggren, P. (1997). Sponge carrying by Indian Ocean bottlenose dolphins: Possible tool-use by a delphinid. *Ethology, 103*, 454–465.

Suddendorf, T., & Whiten, A. (2001). Mental evolution and development: Evidence for secondary representation in children, great apes, and other animals. *Psychological Bulletin, 127*, 629–650.

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13 month olds. *Psychological Science, 18*, 580–586.

Tamaki, N., Morisaka, T., & Taki, M. (2006). Does body contact contribute towards repairing relationships?: The association between flipper-rubbing and aggressive behavior in captive bottlenose dolphins. *Behavioural Processes, 73*, 209–215.

Tomasello, M. (1995). Joint attention as social cognition. In C. Moore & P. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 103–130). New York, NY: Psychology Press.

Tomonaga, M., Tanaka, M., Matsuzawa, T., Myowa-Yamakoshi, M., Kosugi, D., Mizuno, Y., ... Bard, K. A. (2004). Development of social cognition in infant chimpanzees (*Pan troglodytes*): Face recognition, smiling, gaze, and the lack of triadic interactions. *Japanese Psychological Research, 46*, 227–235.

Tomonaga, M., & Uwano, Y. (2010). Bottlenose dolphins' (*Tursiops truncatus*) theory of mind as demonstrated by responses to their trainers' attentional states. *International Journal of Comparative Psychology, 23*, 386–400.

Tschudin, A. (2001). 'Mind-reading' mammals: Attribution of belief tasks with dolphins. *Animal Welfare, 10*, 119–127.

Tschudin, A. (2006). Belief attribution tasks with dolphins: What social minds can reveal about animal rationality. In S. Hurley & M. Nudds (Eds.), *Rational Animals*? (pp. 413–436). Oxford, UK: Oxford University Press.

Tschudin, A., Call, J., Dunbar, R. I., Harris, G., & van der Elst, C. (2001). Comprehension of signs by dolphins (*Tursiops truncatus*). *Journal of Comparative Psychology, 115*, 100–105.

Uyeyama, R. K., Pack, A. A., & Herman, L. M. (2005). How to develop abstract concepts and understanding of complex relationship in dolphins. *Soundings, 30*, 4–7.

Wasserman, E. (2018). Are there minding machines? *International Journal of Comparative Psychology, 31*, 1–3.

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.

Whissell, C., Abramson, C. I., & Barber, K. R. (2013). The search for cognitive terminology: An analysis of comparative psychology journal titles. *Behavioral Sciences, 3*, 133–142.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining functions of wrong beliefs in young children's understanding of deception. *Cognition, 13*, 103–128.

Woodruff, G., & Premack, D. (1979). Intentional communication in the chimpanzee: The development of deception. *Cognition, 7*, 333–362.

Xitco, M. J., Gory, J. D., & Kuczaj, S. A. (2001). Spontaneous pointing by bottlenose dolphins (*Tursiops truncatus*). *Animal Cognition, 4*, 115–123.

Xitco, M. J., Gory, J. D., & Kuczaj, S. A. (2004). Dolphin pointing is linked to the attentional behavior of a receiver. *Animal Cognition, 7*, 231–238.

Zentall, T. (2018). The value of research in comparative cognition. *International Journal of Comparative Psychology, 31*, 1–5.

Financial conflict of interest: No stated conflicts.
Conflict of interest: No stated conflicts.

Submitted: *February 1, 2018*
Revision submitted: *February 1, 2018*
Accepted: *April 17th, 2018*