DISCUSSION TWO

Prior to beginning work on this discussion, please read the following required sources:

- "Behaviorism and Essential Concepts for Pediatric Behavioral Sleep Medicine"
- "When a Clear Strong Voice Was Needed: A Retrospective Review of Watson's (1924/1930) Behaviorism"
- "A Preliminary Application of Social Cognitive Theory to Nonsuicidal Self-Injury"
- "Teaching Stimulus Control Via Class-Wide Multiple Schedules of Reinforcement in Public Elementary School Classrooms"
- "Supporting Students with Autism Spectrum Disorder: Implementing Applied Behavior Analysis Strategies Within the Classroom Setting"

In your initial post, include the following:

- Briefly summarize the foundational approaches of behaviorism and social learning theories and the early theorists associated with them.
- Describe how those two theories might be utilized in practice in the career path that interests you.
- Evaluate any issues and cultural considerations associated with your assigned theorist.
- Analyze and describe how the APA's Ethical Principles of Psychologists and Code of Conduct might affect the implementation of your theorist's personality assessments.
- Assess the types of personality measurements and research designs associated with your assigned theorist, and describe how they have evolved.

Your initial post should be a minimum of 500 words.

Behaviorism and Essential Concepts for Pediatric Behavioral Sleep Medicine

3

Pediatric behavioral sleep medicine has its roots in behaviorism, a set of theories that help to explain how a behavior develops, how this behavior is reinforced, and what must be done for the behavior to change or cease. Thus, the focus of this chapter is to provide readers a basic review of behaviorism, as well as an introduction to the essential concepts in pediatric behavioral sleep medicine that are related to behaviorism. A sample of texts that provide a comprehensive review of behaviorism can be found in the Resources section.

The Role of Behaviorism in Behavioral Sleep Medicine

Behaviorism is a school of psychology dedicated to the study of objective behavior. A key feature of behaviorism is that behavior is not only observable but also can be developed

http://dx.doi.org/10.1037/14645-004
Pediatric Sleep Problems: A Clinician's Guide to Behavioral Interventions, by L. J. Meltzer and V. M. Crabtree

and changed with conditioning. The two major types of conditioning are *classical* and *operant*. Both are relevant for the treatment of pediatric behavioral sleep problems.

CLASSICAL CONDITIONING

Classical conditioning is one way to explain observable behavior change, where interactions with the environment are observed to shape behavior. Early studies of classical conditioning were conducted by Ivan Pavlov, who observed that when he paired a neutral stimulus (bell) with an environmental stimulus (food that causes the dog to salivate), a learned response (salivate when hearing a bell) occurred. In this example, the unconditioned stimulus (the sight or smell of the food) creates an unconditioned or automatic response (salivating). Pavlov then repeatedly paired a neutral stimulus (bell) that previously did not elicit the response (salivating) with the unconditioned stimulus (food). Over time, even in the absence of food, the dogs would salivate upon hearing the bell, demonstrating a conditioned response to the bell.

In classical conditioning, once a behavior has developed, it can be changed or unconditioned through extinction. *Extinction* occurs when the conditioned stimulus is repeatedly presented in the absence of the unconditioned stimulus. In the case of Pavlov, the conditioned response of salivating to the sound of the bell could be extinguished if the bell is repeatedly sounded without the presentation of food. Classical conditioning helps us to understand both the development and treatment of certain pediatric sleep problems.

Classical Conditioning and Enuresis

One of the first treatments in behavioral sleep medicine used classical conditioning for the treatment of enuresis. The bell-and-pad treatment was developed by Mowrer and published in 1938. From a behavior theory standpoint, the bell serves as an unconditioned stimulus, waking the child when a moisture sensor (the pad) becomes wet. Thus, pairing bladder distension (conditioned stimulus) with the sound of the bell (unconditioned stimulus) will result in waking to urinate (conditioned response). As described in Chapter 12 of this volume, more than 75 years later, variations of the bell and pad remain the most effective long-term treatment for enuresis.

Classical Conditioning and Sleep-Onset Associations

Night wakings, one of the most common presenting problems for young children, are most commonly due to negative sleep-onset associations.

Under the framework of classical conditioning, a *sleep-onset association* is a condition repeatedly paired with falling asleep, such that falling asleep becomes linked to that condition. Because nighttime arousals are typical (with all children waking two to six times per night), any sleep-onset association present at bedtime will also need to be present following typical nighttime arousals, helping the child return to sleep.

There are two types of sleep-onset associations. First, a *positive sleep-onset association* is a condition the child can create for himself (e.g., sucking his thumb, cuddling with a favorite stuffed animal or blanket). With a positive sleep-onset association, a child is able to put himself back to sleep following typical nighttime arousals. Infants who are able return to sleep independently (a developmental skill that occurs between 3 and 6 months of age for healthy, typically developing children) are often called *self-soothers*.

A *negative sleep-onset association* is a condition that requires someone or something else to facilitate sleep. This can include parental assistance (e.g., rocking an infant to sleep, lying with a school-age child), environmental stimuli (e.g., the sound of a vacuum cleaner), or a location other than the child's typical sleeping environment (e.g., pushed in the stroller, in the parents' bed). In the case of a negative sleep-onset association, the child pairs the skill of falling asleep with the external support. When seeking treatment, many parents do not complain about having to rock their child to sleep or lie with him until he is asleep at bedtime, but rather the concern is the need for this support every 2 hours following typical nighttime arousals (i.e., rocking the infant back to sleep every 2 hours, the child migrating to his parents' bed during the night). Infants with negative sleep-onset associations are sometimes called *signalers*, as they will cry following typical nighttime arousals because of their inability to return to sleep without assistance.

Because of the powerful role of sleep-onset associations in facilitating sleep onset and sleep maintenance, it is essential to identify negative sleep-onset associations (i.e., "that thing" that helps children fall asleep at bedtime or return to sleep following nighttime arousals). More information about evaluating for sleep-onset associations is provided in Chapter 4. Interventions to remove sleep-onset associations are described in Chapters 6 and 7. These include both rapid and gradual approaches to extinguish the association between sleep and a negative sleep-onset association (Honaker & Meltzer, 2014; Mindell et al., 2006; Morgenthaler, Owens, et al., 2006).

Classical Conditioning and Insomnia

Difficulties initiating and maintaining sleep are the hallmark features of insomnia. Whereas in younger children this is commonly due to a negative sleep-onset association, in older children and adolescents, insomnia

looks more like that of adults. In other words, youth can develop a learned association between their bed and hyperarousal. What this looks like in practice is a patient may feel incredibly sleepy at bedtime, but as soon as she gets into bed, she becomes hyperaroused. Stimulus control therapy (SCT) is one of the primary components of treatment for insomnia (Bootzin, 1972; Morgenthaler, Kramer, et al., 2006). As described in Chapter 9, the goal of SCT is to extinguish this conditioned association by having the patient get out of bed if she is not asleep in 15 to 20 minutes, only returning to bed again when she feels sleepy. This helps weaken the association between lying in bed and not sleeping. Sometimes this process needs to be repeated multiple times in one night. When repeated every night for multiple nights, the association between sleep and bed is strengthened, and the patient's body is eventually retrained to fall asleep quickly once she is in bed.

OPERANT CONDITIONING

Operant conditioning is another way to explain observable behavioral change. However, unlike classical conditioning, behavior is shaped by consequences rather than antecedents. In other words, behaviors that are reinforced will increase in frequency, and behaviors that are punished will decrease in frequency. Operant conditioning is seen in everyday life. For example, when many parents potty train their young children, they will use a treat (a preferred sticker or piece of candy) to reward the child for using the potty. The treat then serves as positive reinforcement for using the potty, making the child more likely to sit on the potty again. In this case, the treat is an intentional reinforcer. However, accidental reinforcement also occurs commonly when reinforcers are provided following behaviors one does not want to increase. For example, most young children want attention from their parents more than anything else and will often do anything (including misbehave) to get their parent's attention. So when a child misbehaves and is subsequently yelled at by his mother, the child's misbehavior is being accidentally reinforced by the attention of his mother, so the unwanted behavior is likely to continue. However, if the mother provides differential attention, consistently ignoring the unwanted behavior but praises him for sharing toys and playing nicely with his sister, over time the unwanted behavior will decrease because it is no longer accessing attention, and the positive behaviors that access attention will increase.

There are two kinds of reinforcement that shape behavior. *Positive reinforcement* increases the likelihood a behavior will occur by adding something after the desired behavior (e.g., praise for the child who shares toys with his sister, stickers for the child who brushes her teeth). *Nega-*

tive reinforcement increases the likelihood a behavior will occur by stopping or removing a negative stimulus. For example, a child who does not want to eat his broccoli throws a tantrum at dinner until his parents remove the broccoli from his plate. Because his parents always give in to the tantrums by removing the broccoli, he will scream and fuss any time they put broccoli on his plate. Thus, his tantrums are being negatively reinforced as the stimulus (broccoli) is repeatedly removed.

In operant conditioning, behavior is changed with two primary reinforcement schedules. *Continuous reinforcement* is consistent and predictable, such that every instance of behavior contacts reinforcement. For example, a child is praised by her mother every time she hangs up her coat in the closet. *Intermittent reinforcement*, however, is unpredictable and does not occur after every instance of behavior, making it more difficult to extinguish a behavior. For example, at the grocery store a child begs for candy and his mother usually says no. But on a couple of occasions, the mother has given in and bought the candy. Now every time they go to the grocery store, the child will beg for candy in the hopes that this is one of the days his mother will say yes. The principles of operant conditioning are also commonly seen in behavioral treatments for pediatric sleep problems; we discuss these below.

Operant Conditioning and Bedtime Problems

Many parents struggle with limit setting around bedtime because they use an intermittent reinforcement schedule. For example, one night a parent may ignore multiple requests for attention at bedtime (e.g., another drink of water, one more bedtime book), but the next night the parent gives in and reinforces the child's behavior. This can result in bedtime stalling that is ultimately frustrating and exhausting for both the parent and child. However, when positive reinforcement for staying in bed is continuously provided (and no accidental reinforcement of out-of-bed behavior occurs), this bedtime stalling will decrease, as it no longer accesses reinforcement. In other words, if every night a child earns a token for every 10 minutes she stays in bed, for example, and in the morning she can exchange these tokens for a small prize, over time the child will be less likely to stall at bedtime in the hopes of earning more tokens. Another example is an infant who does not want to fall asleep alone, so she cries for 30 minutes until her mother rocks her to sleep. By rocking the child, the mother removes the stimulus of crying and achieves the goal of her baby falling asleep. For the infant, the stimulus of being alone is removed, and on subsequent nights the baby has simply learned to cry for prolonged periods of time to get rocked.

Operant Conditioning and Adherence to Positive Airway Pressure Therapy

As described in Chapter 13, positive airway pressure (PAP) therapy is used to treat obstructive sleep apnea for children who do not respond to or are not candidates for surgical interventions (e.g., removing tonsils and adenoids). Yet PAP therapy is not comfortable and poorly tolerated (a high stream of air pressure is blown into the child's nose and/or mouth to open his airway during sleep). It is common to see a child who becomes combative when a PAP mask is placed on his face, so the parents remove the mask immediately. Thus, the child learns that to remove the stimulus of the mask he needs to hit and bite, increasing the frequency of these unwanted and challenging behaviors in the future.

The Bottom Line

The etiology, maintenance, and treatment of behavioral sleep issues rely heavily on the concepts of both classical and operant conditioning. Whereas the treatment chapters in Part III of this volume also include some treatments based in cognitive theories (e.g., addressing dysfunctional thoughts and beliefs about sleep), this chapter has provided basic information on behavioral factors that influence pediatric sleep problems. The concepts of behaviorism are essential to fully appreciate the factors underlying most sleep pediatric sleep problems, as well as to choose the most effective treatment approach when working with a patient/family.

J Youth Adolescence (2016) 45:1560–1574
DOI 10.1007/s10964-016-0449-7

EMPIRICAL RESEARCH

A Preliminary Application of Social Cognitive Theory to Nonsuicidal Self-Injury

Penelope Hasking[1] · **Alyssa Rose**[1]

Received: 13 January 2016 / Accepted: 19 February 2016 / Published online: 26 February 2016

Abstract Researchers have established a relationship between exposure to nonsuicidal self-injury (NSSI), and increased probability of engaging in the behavior, but few have endeavored to explain the mechanisms underlying the relationship. We drew on Social Cognitive Theory to argue that core cognitions, including NSSI outcome expectancies and self-efficacy expectancies, moderate this relationship. We also explored whether knowledge about NSSI and attitudes toward the behavior played a role in this relationship. A sample of 389 university students (73.1 % female, M age = 20.90, SD = 2.36), completed online questionnaires assessing the constructs of interest. Our findings support the application of Social Cognitive Theory to better understanding NSSI, with clear links between expectancies, self-efficacy and NSSI. Further, these cognitions moderated a number of exposure-NSSI relationships. Implications of these findings for theory, research and intervention are discussed.

Keywords NSSI · Self-injury · Expectancies · Self-efficacy · Attitudes · Knowledge

Introduction

A history of nonsuicidal self-injury (NSSI), the deliberate and socially unacceptable destruction of one's own body tissue without conscious suicidal intent (e.g., cutting or burning the skin, self-battery; Nock et al. 2010), is reported by 17 % of adolescents and 13 % of young adults (Muehlenkamp et al. 2012; Swannell et al. 2014), and is associated with depression, anxiety, and increased risk of suicide (Whitlock et al. 2013). NSSI is primarily used as a method of regulating strong emotions, and as such, the majority of research has focused on intrapersonal factors that serve to initiate and maintain NSSI. In efforts to further identify determinants of the behavior, researchers have recently shifted focus to interpersonal factors, including exposure to NSSI within peer groups, online, and in popular media. Drawing on Social Cognitive Theory we propose that observed relationships between exposure to NSSI in peers and/or media can be explained by a number of cognitive factors, specific to NSSI.

The Role of Peers in NSSI

Emerging adulthood is characterized by exploration of identity in love, work and worldview (Arnett 2000), laying the groundwork for more permanent identity formation in adulthood. Although the influence of peers during adolescence has been extensively noted (e.g., Albert et al. 2013), less work has explored the importance of peers during emerging adulthood. Recent work suggests that peer risk taking is associated with delinquency in samples of emerging adults (Curcio et al. in press), and associations between social norms, peer relations and substance use are evident (Perkins 2002). Similarly, exposure to NSSI in peer groups, including viewing wounds and discussing the behavior, is associated with an increased probability of NSSI engagement (e.g., Hasking et al. 2013; Prinstein et al. 2010), while exposure to NSSI in films and other popular media is related to a history of NSSI and may maintain the behavior (e.g., Lewis and Seko 2016; Radovic and Hasking 2013). Some researchers have observed that young people

✉ Penelope Hasking
penelope.hasking@curtin.edu.au

[1] School of Psychology and Speech Pathology, Curtin University, Perth, Australia

gain the initial idea to self-injure from friends, or from television and films (Deliberto and Nock 2008; Heath et al. 2009; Nixon et al. 2008). However, the bulk of this work is with adolescents, with little consideration of how exposure to NSSI in peer groups might relate to NSSI for emerging adults. Further, little consideration has been given as to how exposure to NSSI might increase the probability of engaging in the behavior or how often it is used.

Social Cognitive Theory

In Social Learning Theory, Bandura (1977) posits that a person's behavior influences, and is influenced by, both the social world and personal characteristics, and that learning occurs in the social context. As such, he would predict that people learn behavior through exposure via others. Importantly, when the person modelling the behavior is viewed as one to be emulated (e.g., a celebrity) or the individual identifies with the model (i.e., by sharing similar characteristics) the chances of adopting the modelled behavior are increased. This may explain the increased chance of NSSI when a person is aware a close friend or celebrity self-injures. Social Cognitive Theory (Bandura 1986, 1989) extends traditional learning theories to emphasize the role of two important cognitive processes.

First, *outcome expectancies*, or the anticipated consequences of one's actions, affect which behaviors people will engage in. Behaviors that are anticipated to result in positive or desirable outcomes are readily adopted, while behaviors linked with anticipated negative outcomes are avoided (Bandura 1989). Functional accounts of NSSI imply that people engage in NSSI expecting a particular outcome (e.g., reduced negative affect; Bentley et al. 2014). Yet, there have been few efforts to understand how NSSI-specific cognitions might play a role in differentiating people who do and do not self-injure. Importantly, outcome expectancies are not only formed from personal experience, but also through vicarious learning, allowing a role for exposure to NSSI in forming NSSI-related cognitions. In related areas, there is a strong history of exploring outcome expectancies; alcohol expectancy theory has been heavily researched for over 35 years (e.g., Brown et al. 1980, 1985; Jones et al. 2001), with results used to inform expectancy challenge interventions (Darkes and Goldman 1998). Similar work specific to NSSI may equally inform prevention and intervention efforts.

Second, Bandura (1989, 1997) conceptualized *self-efficacy expectancies*, the belief an individual has in their ability to succeed in specific situations, as a crucial mechanism of human agency. When presented with a particular situation, Bandura purports that a person makes self-efficacy judgments by considering factors associated with the specific situation (Bandura 1982). General self-efficacy has previously been noted to be lower in people who self-injure (e.g., Nock 2009; Tatnell et al. 2014), however, to the best of our knowledge, no researchers have explored self-efficacy judgements specific to NSSI (e.g., a belief in the ability to resist NSSI). In suicide research, Joiner (1999) proposes that suicidal behavior requires an acquired capability, akin to an increased self-efficacy to act on suicidal ideation. Conversely, lowered self-efficacy to resist suicidal action is related to more severe suicidal ideation and number of suicide attempts (Czyz et al. 2014). Again, borrowing from research in drinking behavior, drinking refusal self-efficacy (i.e., an individual's belief in their ability to refrain from drinking alcohol) has repeatedly been associated with reduced drinking in both clinical and community samples (Lee et al. 1999), and is particularly salient in predicting relapse (Greenfield et al. 2000; Kadden and Litt 2011). Should similar cognitions concerning the ability to resist self-injuring when the urge presents itself be salient in governing NSSI this will provide novel avenues for future work in relapse prevention.

Knowledge and Attitudes

While not explicitly articulated within Social Cognitive Theory, exposure to NSSI in peer groups or media may also inform knowledge about or attitudes towards the behavior. Generally, accurate knowledge is related to more positive attitudes toward NSSI and the people who engage in the behavior (e.g., Berger et al. 2015; Muehlenkamp et al. 2013). Such knowledge might include an understanding of NSSI as primarily affect regulatory rather than attention seeking, or statistics about the nature and extent of the behavior. While not central to Social Cognitive Theory, related theories (e.g., Theory of Planned Behavior; Ajzen 1991) specify an explicit role for attitudes in informing whether someone intends to engage in a given behavior, with more favorable attitudes toward the behavior increasing behavioral intentions. For this reason, we also aim to explore whether knowledge and attitudes moderate the relationship between NSSI exposure and NSSI exhibited by our participants.

Where someone acquires their knowledge about NSSI varies, but given that some people first encounter the idea of NSSI from friends or media (Deliberto and Nock 2008; Heath et al. 2009; Nixon et al. 2008), it seems intuitive that such exposure could be one source of knowledge about the behavior, its functions and its consequences. The level of knowledge someone has regarding NSSI may partially depend on the accuracy of portrayals to which they are exposed. The experience of NSSI observed in peers will be largely idiosyncratic, so knowledge about NSSI might depend on the specific situation to which individuals are exposed. Viewing movies featuring scenes of NSSI appears

related to more accurate knowledge (i.e. less confusion of myths with facts about NSSI), but it is not clear how this knowledge is related to engagement in NSSI (Radovic and Hasking 2013).

Finally, while significant work has been conducted in understanding the attitudes health professionals and school staff hold toward NSSI (e.g., Berger et al. 2015; Heath et al. 2011; Muehlenkamp et al. 2013), less work has examined how attitudes an individual holds about NSSI might relate to probability they engage in NSSI. While permissive attitudes toward suicide differentiate people exhibiting suicidal behavior from those who do not (Nock et al. 2010; O'Connor and Armitage 2003), people who self-injure hold stronger negative implicit attitudes (e.g., cutting is bad) than people who do not self-injure (Nock and Banaji 2007). This might suggest that rather than viewing the behavior favorably, people who self-injure internalize stigmatized views, which could exacerbate distress and facilitate on-going NSSI.

The Current Study

We aimed to explore whether outcome expectancies and self-efficacy (i.e., self-efficacy to resist self-injury), knowledge about NSSI and attitudes towards NSSI help explain the relationship between exposure to NSSI by peers or in movies and the probability of an individual engaging in the behavior. Specifically we asked: Is the relationship between exposure to NSSI and NSSI engagement moderated by outcome expectancies, self-efficacy to resist NSSI, knowledge and attitudes?

Despite a proposed dose–response relationship, in which greater exposure is thought to confer greater probability of NSSI (Schmidtke and Häfner 1988), few researchers have examined the correlation between number of self-injuring peers and NSSI behavior, although number of films, and number of self-injuring characters participants identify with, have each been associated with NSSI (Radovic and Hasking 2013). Our predictors concerning peer exposure will include both knowing someone who self-injures, and how many peers self-injure, while predictors regarding exposure in movies will include number of films seen and number of self-injuring characters participants identify with. Both history of NSSI (yes/no), and frequency of the behavior performed by those who self-injure will be explored as outcomes.

Although we consider this work exploratory, we expect greater exposure to NSSI to increase the probability an individual self-injures and be associated with frequency of the behavior in those who self-injure. Yet, we expect this will be an indirect relationship, moderated by the cognitive variables (expectancies, self-efficacy, knowledge, attitudes). For example, exposure to NSSI may only relate to an increased probability of self-injury if the individual has a lowered ability to resist the urge to self-injure.

Methods

Participants

Participants were 389 undergraduate students, aged between 18–30 years ($M = 20.90$, $SD = 2.36$) recruited from a large Australian university. Of the sample 103 were male (26.6 %), 283 were female (73.1 %), and 1 was transgender (the remainder did not specify). The majority of participants were born in Australia (73.5 %) and lived at home with family[1] (74.3 %) or with house mates (12.3 %). Most of the sample was in either the first or second year of university (76.3 %), and studying full-time (92.3 %).

Materials

Self-Injury

Participants' self-injury history was assessed with Section I of the Inventory of Statements About Self-Injury (Klonsky and Glenn 2009). Participants reporting a history of NSSI responded to questions about age of onset, frequency, and main form of self-injury. The measure has established test-retest reliability in samples of young adults ($r = .85$; Klonsky and Olino 2008).

Exposure to NSSI

To assess NSSI exposure, participants were asked whether any of their friends self-injured (0 = no; 1 = yes), and if yes, how many of their friends self-injured. Exposure to films containing NSSI was assessed using the procedure outlined by Trewavas et al. (2010). The initial list was updated by entering search terms such as "self-injury", "self-harm", "self-mutilation" and "parasuicide" into both Google searches and IMDB (Internet Movie Database). To enhance the probability that students had seen the films, and could identify with characters, films made before 1990, horror, science fiction, animated films, non-English language films and films made solely for television were excluded. The final list comprised 47 feature films. Participants were asked whether they had seen each film (Yes/No), and whether they had identified with the character that self-injured (Yes/No).

[1] It is customary in Australia for students to remain at home rather than travel away to college.

NSSI Expectancies

Given the lack of prior research exploring NSSI-related outcome expectancies we developed our own measure for this study, using items elicited from the literature on NSSI and existing measures of alcohol outcome expectancies (e.g., Lee et al. 2003). Initially the questionnaire comprised 42 items, assessed on a 4-point Likert scale indicating the degree to which participants anticipate the given outcome (e.g., "How likely is it that after self-injuring: I would feel better about myself.").

Principal axis factoring, with oblique rotation, was used to extract factors from the questionnaire items. Eigen values over one and examination of the scree plot were used to determine the number of factors. Items were only included if they loaded unambiguously on a factor (loadings $\geq .40$), were conceptually coherent, and did not cross-load on factors. No item had more than 1.5 % missing data; consequently missing data were replaced with the series mean. Initial analysis produced eight factors with eigen values greater than one, which explained 60.81 % of the variance. However examination of the scree plot suggested a three-factor solution. After removal of two cross-loading items, four items which did not load on any factor and thirteen items with low communalities, the remaining 23 items produced a three-factor solution explaining 55.15 % of the variance.[2] The resultant factors reflect positive expectancies ("I would feel better about myself" $\alpha = .92$), negative affect expectancies ("I would feel ashamed" $\alpha = .90$) and expectancies regarding the negative social consequences of NSSI ("My friends would be disgusted" $\alpha = .76$; Table 1). The first two factors were not correlated ($r = .01$, $p = .90$). Positive ($r = .20$, $p < .001$) and negative ($r = .37$, $p < .001$) expectancies were both correlated with negative social expectancies.

Self-Efficacy to Resist NSSI

Originally developed to measure an individual's belief in their ability to resist suicidal behavior, the 6-item Self-efficacy to Resist Suicidal Action Scale was adapted to measure an individual's belief in their ability to resist self-injuring (Czyz et al. 2014; e.g., "How certain are you that you will not self-injure in the future?"). Answers range from (1) very uncertain to (6) very certain, with higher scores indicating a stronger self-efficacy to resist self-injury. Initial validation revealed strong internal consistency (Cronbach's $\alpha = .96$) and good convergent validity, being highly correlated with severity of suicidal ideation ($r = -.59$, $p < .001$; Czyz et al. 2014). Internal consistency was also strong in the current study (Cronbach's $\alpha = .92$).

Knowledge of NSSI

The 20-item Self-Injury Knowledge Questionnaire (Warm et al. 2003) was used to assess the participants' knowledge of NSSI. This measure contains ten true statements about NSSI (e.g., "Self-injury is a coping strategy") and ten considered to be myths (e.g., "Self-injury is a woman's problem"), which the authors identified through a literature review and input from young people who self-injure. Participants were asked to indicate whether statements were true or false, with the total number of correct responses recorded. In the current sample Cronbach's $\alpha = .64$.

Attitudes to NSSI

The participant's attitudes towards self-injury were assessed with the scale developed by Muehlenkamp et al. (2013) for use with health professionals. The wording of questionnaire items was altered for applicability to the general public. The measure is comprised of 20 items drawn from three pre-existing measures (Muehlenkamp et al. 2013). Participants were required to answer via a 5 point scale ranging from strongly agree to strongly disagree. Scores are summated to produce three subscales: positive empathy (e.g., "I feel concern for people who self-injure"), negative attitudes (e.g., "Those who self-injure are often doing it to manipulate other people"), and perceived knowledge and comfort (e.g., "I feel knowledgeable about the area of self-injury"). The measure reliably differentiates the attitudes of health professionals with varying degrees of experience with NSSI (Muehlenkamp et al. 2013). In the current sample positive empathy $\alpha = .76$; negative attitudes $\alpha = .83$; comfort $\alpha = .82$.

Psychological Distress

To aid interpretation and minimize confounding of cognitions and mental health, symptoms of psychological distress were statistically controlled in all regression analyses. Pre-existing symptoms were assessed with the Depression, Anxiety, Stress Scales (DASS; Lovibond and Lovibond

[2] At the suggestion of an anonymous reviewer we conducted a parallel analysis with 1000 samples (O'Connor 2000), using R-Menu v2 in SPSS (Courtney 2013), to confirm the number of factors for extraction. Initial analysis with all 42 items suggested five factors be extracted. After removal of items with poor loadings and communalities we conducted the analyses again with the remaining 23 items; 6 factors were suggested. Examination of the factor loadings when 6 factors were requested revealed the same three factors as outlined above. The additional three factors were primarily composed of the same items but with negative factor loadings. Consequently we believe the three factors we extracted in the Exploratory Factor Analysis are an accurate representation of the factor structure of our measure.

Table 1 Three factor solution for the NSSI expectancy questionnaire

Item	1[a]	2[b]	3[c]	Communalities
I would feel better about myself	**.82**	−.05	−.12	.67
I would feel happy	**.82**	−.05	−.17	.73
I would feel less frustrated with the world	**.79**	−.05	.10	.70
The future would seem more optimistic	**.79**	−.10	−.14	.61
I would feel relieved	**.78**	−.03	−.12	.66
I would feel calm	**.77**	−.05	−.16	.65
I would feel exhilarated	**.72**	.05	−.11	.53
I would have more confidence	**.70**	−.08	−.10	.58
I would feel I could manage stressful events in the future	**.63**	.05	−.10	.39
I would feel I have successfully achieved something	**.62**	.09	−.09	.38
I would feel alone	.02	**.79**	−.32	.69
I would feel like a failure	.05	**.74**	−.37	.56
I would feel angry	.00	**.73**	−.17	.61
I would feel emotionally drained	.02	**.73**	−.39	.55
I would feel disappointed	−.11	**.71**	−.33	.47
I would feel anxious	−.06	**.69**	−.14	.53
I would feel ashamed	−.17	**.62**	−.38	.39
I would feel restless	.13	**.60**	−.15	.35
I would feel depressed	−.17	**.59**	−.23	.34
My friends would be disgusted	.09	.31	**.83**	.71
My family would be disgusted	.18	.26	**.83**	.68
My parents would be angry	.17	.25	**.80**	.53
My friends would not approve of me	.21	.38	**.70**	.43
Cronbach's alpha	.92	.90	.76	

Items loading on each factor are given in bold

[a] Positive outcome expectancies; [b] negative affect expectancies; [c] social outcome expectancies

1995). The DASS is a 21-item questionnaire assessing symptom severity on a 4-point Likert scale. Psychometric evaluation of the measure demonstrated good internal consistency and validity with undergraduate samples (Depression: $\alpha = .81$; Anxiety: $\alpha = .73$; Stress: $\alpha = .81$; Lovibond and Lovibond 1995) and also sound internal consistency in the current sample (Depression: $\alpha = .91$; Anxiety: $\alpha = .86$; Stress: $\alpha = .88$).

Procedure

After obtaining approval from the Human Research Ethics Committee, participants were recruited through advertisements on an online booking system for undergraduate psychology students interested in participating in research for course credit. The advertisement provided a definition of NSSI, and informed that participants would be asked to respond to questions regarding experiences they have had with self-injury (if any), what they think might happen if someone self-injures, their own emotional health and levels of distress. The advertisement explicitly stated that we required both participants who did, and who did not, self-injure.

Interested participants were directed to the online survey, which could be completed in their own time at a place of their choosing (approx. 30–45 min). All participants were fully informed of the nature of the study, confidentiality, data storage arrangements and of their right to withdraw at any point during questionnaire completion. A list of useful resources (i.e., on-campus and telephone counselling, websites) and an electronic booklet about NSSI were provided to all participants.

Data Analysis

A series of eight moderated multiple regressions were conducted in SPSS v23. The first two analyses assessed the relationships between peer exposure (1. NSSI in peers, scored yes/no; 2. Number of peers who self-injure) and probability of NSSI, with logistic regressions. In the third and fourth analyses the relationships between these variables and frequency of NSSI, in participants who self-injured, were assessed with linear regressions. The relationships between exposure to movies (1. Number of movies seen, 2. Number of characters identified with) and probability of NSSI (yes/no) were similarly assessed with

logistic regression. Finally the relationships between movie exposure and frequency of NSSI in the sub-sample that self-injured were assessed with linear regressions. Prior to interpretation of the regression results, the relevant logistic and linear regression assumption tests and checks were conducted, all of which were met. Multiple imputation was used to impute no more than 5 % on the continuous variables (ranging from 2.4 % on self-efficacy to 4.98 % on depression).

In all analyses symptoms of depression, anxiety and stress were statistically controlled by entering them in the first step. The predictor variable was entered in step two, and expectancies, self-efficacy, knowledge of NSSI, and attitudes were simultaneously entered in the third step. All two-way interaction terms were entered in the final step of each regression. All predictors were mean-centered for analysis and significant interactions probed using simple slopes analysis at ± one standard deviation from the mean (Aiken and West 1991). Given the number of analyses significance was estimated at $\alpha < .01$.[3]

Results

Preliminary Analyses

Of the sample 114 (29.5 %) had engaged in NSSI, with a mean age of onset of 14.60 years ($SD = 2.14$). Among self-injurers, 42.5 % had self-injured in the last 12 months, and of these 31.3 % had self-injured on five or more occasions. Cutting was endorsed as the primary means of NSSI for 47.7 % of the sample, and self-battery by 14.7 %, but the majority (77.6 %) engaged more than one method of NSSI. Most participants reported knowing at least one friend who self-injured ($n = 266$; 69.1 %), with an average of 2.92 self-injuring friends ($SD = 2.83$). The majority of people who know someone who self-injured (90.2 %) reported their friend had self-injured in the last year, primarily by cutting (84.2 %). Most participants had seen at least one movie featuring a scene of NSSI (91.3 %), with a mean of 5.31 ($SD = 4.28$) movies. Exactly half the participants (50.0 %) identified with at least one character ($M = 2.02$, $SD = 1.37$) who self-injured.

There was no gender difference in history of NSSI, χ^2 (1, $N = 384$) = 2.54, $p = .11$, but female participants were more likely to know someone who self-injured, χ^2 (1, $N = 384$) = 10.82, $p < .001$, Cramer's V = .17. There were no gender differences in the number of movies seen or characters participants identified with ($p > .05$). People who knew someone who self-injured were more likely to report a history of NSSI, χ^2 (1, $N = 384$) = 12.61, $p < .001$, Cramer's V = .18. Neither engaging in NSSI, nor knowing someone who self-injured, were related to how many movies participants had seen or the number of characters identified with (all $p > .05$).

As seen in Table 2, participants who self-injured reported more depression, anxiety and stress than participants with no history of NSSI. These participants also reported stronger positive expectancies, weaker negative expectancies, and reduced self-efficacy to resist self-injury. Participants who self-injured also reported a more positive attitude and more comfort with NSSI than participants with no NSSI history (Table 2).

Peer Exposure

Knowing a Friend Who Self-Injures

The logistic regression predicting probability of NSSI, when knowing a friend who self-injures was the predictor, significantly differentiated participants who self-injured and those who did not, accounting for between 31.9 % (Cox & Snell R^2) and 45.3 % (Nagelkerke R^2) variance. At Step 1, only symptoms of depression were related to probability of NSSI. At Step 2, knowing a friend who self-injures was positively related to probability of NSSI reported by participants, after controlling for psychological distress (Table 3). The cognitive variables significantly contributed to the regression equation, with negative expectancy and self-efficacy to resist NSSI inversely related to probability of NSSI, and social expectancies and more accurate knowledge increasing the probability the participants self-injured. None of the interaction terms significantly predicted probability of NSSI.

The linear regression predicting frequency of NSSI in the sub-sample who self-injured, when knowing someone who self-injured was the predictor, accounted for significant variance in frequency of the behavior, $R^2 = .16$, $p = .001$. Knowing a friend who self-injured was not directly related to frequency of NSSI, but self-efficacy to resist was inversely related to frequency. In the final step, self-efficacy also moderated the relationship between knowing a friend self-injured and frequency; a negative relationship was observed when self-efficacy to resist NSSI was high, $b = -.32$, $t = -3.09$, $p = .001$, but there was no relationship when self-efficacy was low, $b = .16$, $t = 1.40$, $p = .08$ (see Fig. 1).

Number of Friends Who Self-Injure

The logistic regression exploring the relationship between number of friends known to self-injure and probability of

[3] We acknowledge that the choice of alpha cut-off is completely arbitrary. We set this level to reduce Type I error but also reduce the risk of Type II error likely with more conservative (e.g., Bonferonni) adjustments (Perneger 1998; Rothman 1990).

Table 2 Differences between participants who self-injured and those who did not on key variables

NSSI history	Yes ($n = 119$)		No ($n = 302$)		t	p
	M	*SD*	*M*	*SD*		
Depression	15.26	5.36	12.12	4.00	9.07	.000
Anxiety	14.72	5.08	12.56	4.07	7.20	.000
Stress	16.21	4.82	13.67	4.07	7.57	.000
Peer exposure (number)	4.09	4.14	2.31	1.48	5.11	.000
Movies seen	4.90	4.13	4.81	4.42	.53	.88
Characters identified	2.20	1.75	2.27	1.70	.23	.82
Positive expectancies	34.67	8.59	29.55	8.82	4.98	.000
Negative expectancies	40.14	9.26	44.49	7.86	4.57	.000
Social expectancies	13.43	3.31	12.98	3.12	1.22	.22
Self-efficacy	23.65	8.57	29.00	7.11	6.26	.000
Knowledge	10.76	1.59	10.61	1.64	.83	.41
Positive attitudes	29.54	3.90	28.40	3.51	2.76	.006
Negative attitudes	11.72	3.62	12.38	3.68	1.60	.11
Comfort	23.92	5.00	21.06	4.42	5.49	.000

Table 3 The relationship between knowing a friend who self-injures and both probability of NSSI (full sample) and frequency of NSSI (NSSI sub-sample)

	Probability of NSSI[a]				Frequency of NSSI			
	B	SE *B*	Exp (*B*)	95 % CI	*B*	SE *B*	t	95 % CI
Step one	$\chi^2 = 67.30$, $p < .001$				$R^2 = .06$, $F = 15.36$, $p < .001$			
Depression	.13	.03	1.14***	1.07, 1.21	.04	.01	3.52***	.02, .06
Anxiety	.01	.03	1.01	.95, 1.08	.03	.01	2.03	.00, .05
Stress	.02	.04	1.02	.94, 1.10	−.02	.02	−1.10	−.05, .01
Step two	$\chi^2 = 18.04$, $p < .001$				$\Delta R^2 = .00$, $\Delta F = .47$, $p = .49$			
Friend NSSI[b]	.91	.22	2.48**	1.60, 3.85	−.05	.07	−.69	−.19, .09
Step three	$\chi^2 = 151.18$, $p < .001$				$\Delta R^2 = .07$, $\Delta F = 7.41$, $p < .001$			
Pos. expectancies	.03	.01	1.03	1.01, 1.06	−.00	.00	−.33	−.01, .01
Neg. expectancies	−.12	.02	.89***	.86, .92	.00	.00	.87	−.01, .01
Social expectancies	.14	.04	1.16***	1.07, 1.25	−.00	.01	−.26	−.04, −.02
Self-efficacy	−.08	.02	.92***	.89, .95	−.03	.01	−6.92***	−.04, −.02
Knowledge	.27	.07	1.30***	1.13, 1.50	.01	.02	.27	−.04, .05
Pos. attitudes	.09	.04	1.09	1.01, 1.19	.01	.01	.47	−.02, .03
Neg. attitudes	−.06	.03	.94	.88, .99	−.01	.01	−1.41	−.03, .01
Comfort	.07	.03	1.07	1.02, 1.14	.02	.01	2.11	.00, .03
Step 4[c]	$\chi^2 = 15.64$, $p = .05$				$\Delta R^2 = .03$, $\Delta F = 3.51$, $p = .001$			
Friend NSSI × self-efficacy					−.03	.01	−3.10**	−.05, −.01

** $p < .01$; *** $p < .001$

[a] History of NSSI (1 = yes, 2 = no); [b] knowing a friend who self-injures (yes/no); [c] only significant interactions included in the table

NSSI significantly differentiated those who self-injured and those who did not, explaining between 39.0 % (Cox & Snell R^2) and 53.7 % variance (Nagelkere R^2). After controlling for psychological distress, the number of friends known to self-injure was positively related to a history of NSSI (Table 4). At step 3, the cognitive variables again contributed significant variance in the logistic regression with positive expectancies, social expectancies and accurate knowledge increasing probability of NSSI and negative expectancies and self-efficacy to resist NSSI decreasing the probability of NSSI. At Step 4, none of the interaction terms contributed to the regression.

While the linear regression, with number of friends as the predictor and frequency of NSSI as the criterion,

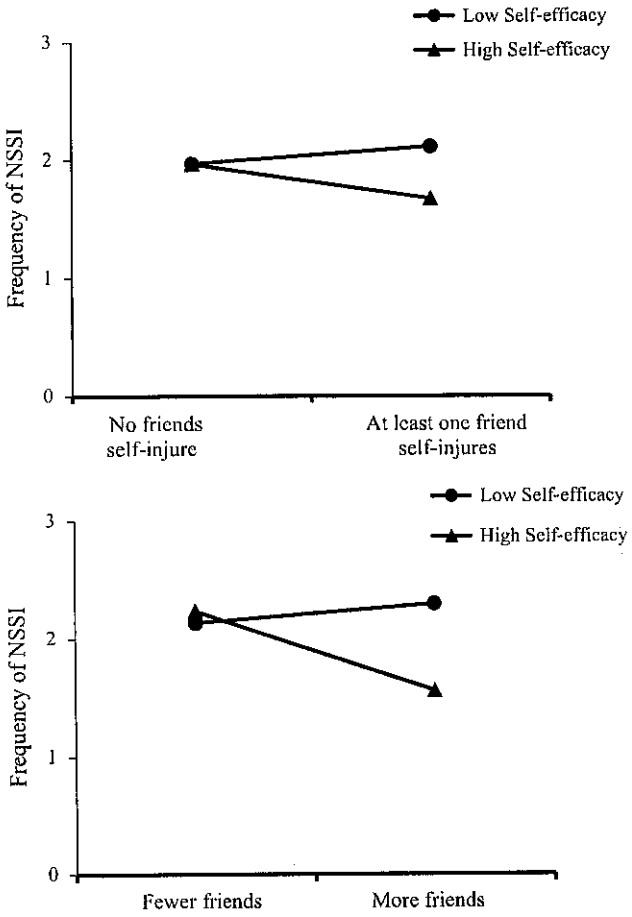


Fig. 1 Self-efficacy moderates the relationship between simply knowing a friend self-injures (*top panel*) and how many friends are known to self-injure (*bottom panel*) and the frequency of NSSI

accounted for significant variance in the frequency of NSSI, $R^2 = .19$, $p < .001$, number of friends known to self-injure was negatively related while self-efficacy to resist NSSI was positively related to the behavior. These counterintuitive relationships might be explained by the observation that, in Step 4 of the analysis, self-efficacy moderated the relationship between number of friends and frequency of NSSI. As seen in Fig. 1, no relationship between number of friends and frequency was observed for participants with less ability to resist NSSI, $b = .04$, $t = 1.52$, $p = .06$, but a negative relationship was observed for those with higher levels of self-efficacy, $b = -.14$, $t = -5.72$, $p < .001$.

Movie Exposure

Number of Movies Seen

When predicting probability of NSSI from the number of movies participants had seen, the logistic regression accounted for between 31.1 % (Coz & Snell R^2) and 44.7 % (Nagelkere R^2) variance. However, the number of movies seen was not directly related to probability of NSSI after controlling for psychological distress (Table 5). At step 3, negative expectancies and self-efficacy to resist NSSI were inversely related to probability of NSSI, while social expectancies and accurate knowledge increased the probability of NSSI. In the final step, self-efficacy to resist NSSI moderated the relationship between the number of movies seen and probability of NSSI. As seen in Fig. 2, a weak positive relationship was observed when self-efficacy was low, $b = .07$, $Z = 1.87$, $p = .06$, while a weak negative relationship was observed when self-efficacy was high, $b = -.08$, $Z = -1.74$, $p = .08$. When predicting frequency of NSSI, from the number of movies seen, the final step of the linear regression, with all variables entered, did not explain a significant amount of variance in the behavior $R^2 = .13$, $p = .32$; self-efficacy emerged as the only significant predictor of frequency of NSSI.

Identification with Self-Injuring Characters

Between 39.8 % (Cox and Snell R^2) and 53.7 % (Nagelkerke R^2) of variance in probability of NSSI was significantly explained by the logistic regression with number of characters with which participants identified entered as the predictor. However, after accounting for psychological distress, the number of characters was not directly related to probability of NSSI; negative expectancies were the only variable to contribute a unique effect (Table 5). However, at Step 4, the relationship between number of characters with which participants identified and probability of NSSI was moderated by both negative and positive expectancies (see Fig. 3). Although the significant interaction implies the individual slopes are different from each other, simple slopes analysis suggests no relationship between number of characters and probability of NSSI for participants with weak, $b = -.03$, $Z = -.18$, $p = .86$, or strong, $b = -.27$, $Z = -1.39$, $p = .16$, negative expectancies. Similarly, there is no relationship for participants with either weak, $b = -.34$, $Z = -1.62$, $p = .10$, or strong, $b = .03$, $Z = .16$, $p = .87$, social expectancies. The inflated error estimates and diminishing power which arise at each step of the regression analysis may account for the lack of significant slopes (Aguinis and Stone 1997).

Finally, the linear regression predicting frequency of NSSI in the self-injuring sub-sample explained a significant amount of variance in the behavior, $R^2 = .17$, $p < .001$, yet after controlling psychological distress, the number of characters with which participants identified was not directly related to NSSI frequency. At step 3, self-efficacy to resist NSSI was again inversely related to NSSI

Table 4 The relationship between number of friends known to self-injure and both probability of NSSI (full sample) and frequency of NSSI (NSSI sub-sample)

	Probability of NSSI[a]				Frequency of NSSI			
	B	SE *B*	Exp (*B*)	95 % CI	*B*	SE *B*	t	95 % CI
Step one	Step $\chi^2 = 38.99$, $p < .001$				$R^2 = .06$, $F = 15.36$, $p < .001$			
Depression	.13	.03	1.14***	1.07, 1.21	.04	.01	3.52***	.02, .06
Anxiety	.01	.03	1.01	.95, 1.08	.03	.01	2.03	.00, .05
Stress	.02	.04	1.02	.94, 1.10	−.02	.02	−1.10	−.05, .01
Step two	Step $\chi^2 = 35.50$, $p < .001$				$\Delta R^2 = .02$, $\Delta F = 13.26$, $p < .001$			
No. friends[b]	.28	.06	1.33***	1.19, 1.48	−.05	.01	−3.64***	−.08, −.02
Step three	Step $\chi^2 = 147.65$, $p < .001$				$\Delta R^2 = .07$, $\Delta F = 7.63$, $p < .001$			
Pos. expectancies	.05	.02	1.05**	1.02, 1.09	.00	.00	.10	−.01, .01
Neg. expectancies	−.14	.02	.87***	.84, .91	.00	.00	.92	−.01, .01
Social expectancies	.16	.05	1.17***	1.06, 1.29	−.03	.01	.06	−.02, .02
Self-efficacy	−.10	.02	.91***	.87, .94	−.03	.01	−7.02***	−.04, −.02
Knowledge	.37	.10	1.44***	1.19, 1.76	.01	.02	.45	−.03, .05
Pos. attitudes	.11	.05	1.12	1.00, 1.24	.01	.01	.79	−.14, .03
Neg. attitudes	−.10	.04	.91	.84, .98	−.01	.01	−1.01	−.03, .01
Comfort	.12	.04	1.13***	1.05, .121	.02	.01	1.81	−.00, .03
Step Four[c]	Step $\chi^2 = 4.88$, $p = .77$				$\Delta R^2 = .05$, $\Delta F = 5.41$, $p < .001$			
No. friends × self-efficacy					−.01	.01	−5.61**	−.05, −.01

** $p < .01$; *** $p < .001$

[a] History of NSSI (1 = yes, 2 = no); [b] how many friends self-injure; [c] only significant interactions included in the table

frequency (Table 6). At step 4, knowledge regarding NSSI moderated the relationship between number of characters and NSSI frequency, as did positive attitudes. However, simple slopes analysis again failed to reveal significant relationships for any individual slope (all $p > .05$, Fig. 4).

Discussion

NSSI is a complex behavior, engaged in by approximately 13 % of community-based young adults (Swannell et al. 2014). With prospective research suggesting a link between NSSI and adverse psychological outcomes, including increased risk of suicide (e.g., Whitlock et al. 2013), understanding factors that serve to initiate and maintain the behavior is paramount. Current research suggests that exposure to NSSI among peers and in popular media is associated with engaging in NSSI (e.g., Hasking et al. 2013; Lewis and Seko 2016; Prinstein et al. 2010); yet to date, the mechanisms underlying this relationship remain elusive. Social Cognitive Theory proposes that the relationship between environmental stimuli and behavior can be explained by individual cognitions regarding the anticipated outcomes of a given behavior (outcome expectancies) and the perceived ability to perform the behavior (self-efficacy expectancies). It is also possible that exposure to NSSI informs knowledge and attitudes regarding the behavior, which could impact on someone's decision to self-injure.

Thus, drawing on Social Cognitive Theory, we aimed to explore whether outcome expectancies and self-efficacy (i.e., self-efficacy to resist self-injury), knowledge about NSSI, and attitudes towards NSSI help explain the relationship between exposure to NSSI and the probability of an individual engaging in the behavior. Overall our findings support the application of Social Cognitive Theory to better understanding NSSI, with consistent links between expectancies, self-efficacy, and NSSI. Further, the key cognitive constructs moderated several of the exposure-outcome relationships. Although preliminary, we suggest our findings provide early evidence of the possible mechanisms of action in the relationship between exposure to NSSI and NSSI outcomes.

Key Findings

Our analyses suggest an elevated rate of NSSI in our sample, relative to other samples of youth (e.g., Swannell et al. 2014). While this may be an artefact of self-selection into a study about NSSI, prior work has also suggested elevated rates in university students (e.g., Hasking et al. 2008; Kokaliari 2014). Heightened distress is commonly noted among university students (Eisenberg et al. 2013), of which NSSI may be one symptom.

Table 5 The relationship between number of movies viewed and both probability of NSSI (full sample) and frequency of NSSI (NSSI subsample)

	Probability of NSSI[a]				Frequency of NSSI			
	B	SE *B*	Exp (*B*)	95 % CI	*B*	SE *B*	t	95 % CI
Step one	Step $\chi^2 = 74.71$, $p < .001$				$R^2 = .06$, $F = 15.24$, $p < .001$			
Depression	.13	.04	1.14***	1.06, 1.22	.04	.01	3.51***	.02, .06
Anxiety	.01	.04	.79	.94, 1.09	.03	.01	2.02	.00, .05
Stress	.06	.05	.20	.97, .16	−.02	.02	−1.10	−.05, .01
Step two	Step $\chi^2 = .04$, $p = .84$				$\Delta R^2 = .00$, $\Delta F = .18$, $p = .67$			
No. movies[b]	−.01	.02	.99	.95, 1.04	−.00	.01	−.43	−.02, .01
Step three	Step $\chi^2 = 104.83$, $p < .001$				$\Delta R^2 = .07$, $\Delta F = 7.07$, $p < .001$			
Pos. expectancies	.02	.01	1.02	.99, 1.05	−.00	.00	−.26	−.01, .02
Neg. expectancies	−.09	.02	.91***	.88, .94	.00	.00	.84	−.01, .01
Social expectancies	.14	.05	1.15***	1.5, 1.05	−.00	.01	−.23	−.03, .02
Self-efficacy	−.09	.02	.92***	.89, .95	−.03	.01	−6.83***	−.04, −.02
Knowledge	.26	.08	1.30**	1.11, 1.52	.01	.01	.52	−.03, .05
Pos. attitudes	.10	.05	1.11	1.00, 1.22	.01	.01	.42	−.02, .03
Neg. attitudes	−.07	.04	.94	.87, 1.00	−.01	.01	−1.21	−.03, .01
Comfort	.08	.03	1.09**	1.02, 1.16	.02	.01	1.90	−.00, .03
Step four[c]	Step $\chi^2 = 21.54$, $p < .01$				$\Delta R^2 = .01$, $\Delta F = 1.16$, $p = .32$			
No. movies × self-efficacy	−.02	.01	.98**	.97, .10				

** $p < .01$; *** $p < .001$

[a] History of NSSI (1 = yes, 2 = no); [b] number of movies viewed; [c] only significant interactions included in the table

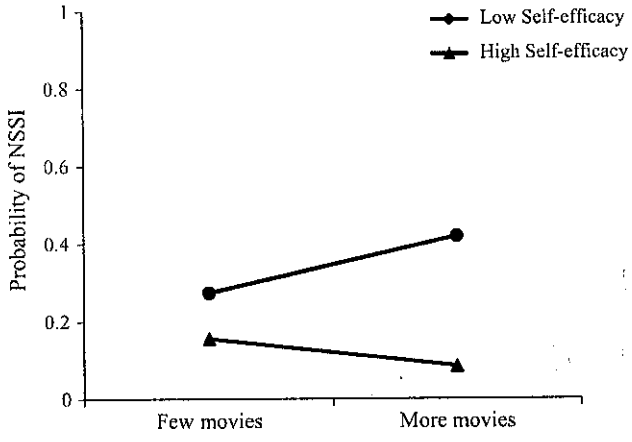


Fig. 2 Self-efficacy moderates the relationship between number of movies viewed and probability of NSSI

However, we also used a detailed checklist to assess NSSI, which is known to produce higher rates than single items or interviews (Swannell et al. 2014). That most participants knew someone who self-injured could also reflect self-selection. Participants in our study recalled seeing fewer films, but identified with more characters than participants in an earlier study which used the same methodology (Radovic and Hasking 2013). One could argue that the high rate of NSSI might explain the identification with movie characters, but these were unrelated in the current study. Further work is needed to fully explore what it means to identify with a movie character and the implications this has for NSSI. Finally, participants who self-injured reported more psychological distress than participants who did not self-injure. This is not an unexpected finding given the association between NSSI and psychological morbidity (e.g., Wilcox et al. 2012). While we controlled these variables in all analyses, we acknowledge that underlying psychological distress may impact the relationships we observed.

Our results support previous research showing awareness of NSSI among peers is associated with self-injurious behavior in the individual (e.g., Prinstein et al. 2010), and suggest a dose–response relationship in conferring risk of NSSI; that is, the more peers known to self-injure the greater the probability that someone self-injures. As noted above, little work has explored the impact of peer exposure on NSSI in emerging adults. It is possible that if the ability to negotiate peer pressure is not established by the end of adolescence, individuals remain susceptible to peer influence throughout their youth, increasing the probability of NSSI if a peer also self-injures (Steinberg and Monahan 2007). We neglected to ask the age at which participants were aware of a peer self-injuring so we are not able to determine whether knowing someone who self-injures increases the probability of NSSI, or whether people who self-injure gravitate toward like others.

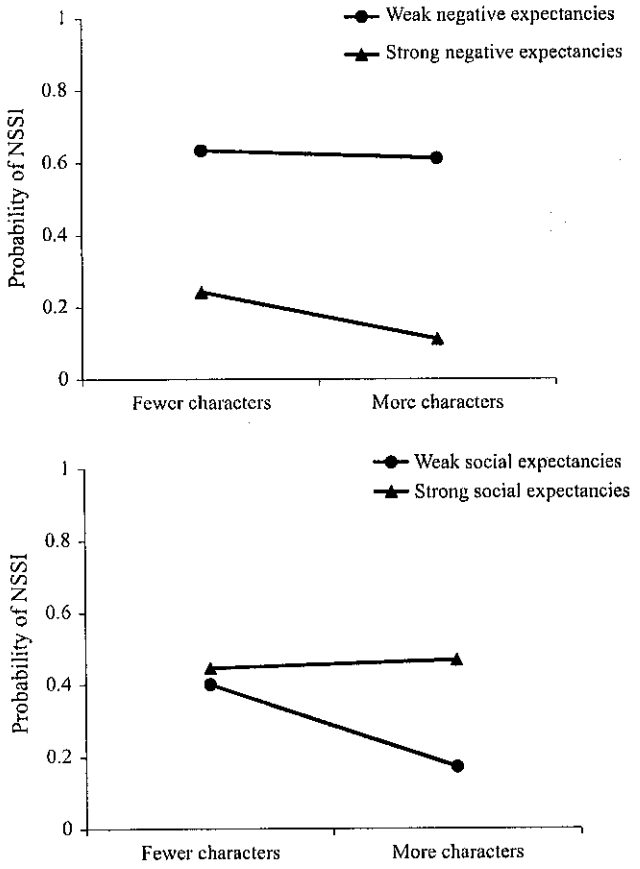


Fig. 3 Negative (*top panel*) and social (*bottom panel*) outcome expectancies moderate the relationship between the number of self-injuring characters the participants identify with and probability of NSSI

The number of friends known to self-injure was inversely related to frequency of the behavior in the sub-sample of self-injurers. At least two hypotheses can be proposed to explain this relationship: (1) Greater exposure to NSSI may increase awareness of the negative consequences of the behavior, which could constrain how often someone self-injures; (2) Isolated individuals who know fewer people in general, as well as fewer who self-injure, may self-injure more frequently as a means of regulating emotion. We also observed that self-efficacy moderated the relationship between peer exposure and frequency of NSSI, such that a high self-efficacy to resist self-injury countered the effect of peer exposure, thus explaining an inverse direct effect (see Fig. 1). However, further work is required to establish the relationship between NSSI in peer groups during this critical developmental period.

The overall lack of influence of NSSI depicted in films goes against populist suggestions that NSSI in popular media might increase the probability of engaging in NSSI. Yet, a number of studies now argue that depictions of NSSI in film and in other media can have both positive and negative effects (e.g., Lewis and Seko 2016; Radovic and Hasking 2013; Swannell et al. 2010), potentially masking any direct relationships. Further, although exposure to movies was not directly related to NSSI the relationship was moderated by self-efficacy and expectancies. Specifically, viewing more movies featuring NSSI appeared to exacerbate the risk conferred by a lowered ability to resist NSSI. Yet, given the cross-sectional nature of the data it is not clear whether increased exposure to NSSI reduces self-efficacy or whether people with a lowered self-efficacy seek out such movies. Identification with movie characters also appeared to interact with the cognitive variables, although individual slopes were not statistically significant. Generally, it seems how people think about NSSI, and the outcomes they anticipate, are more salient predictors of behavior than identifying with fictional movie characters.

It is interesting to note that participants scored an average of only 50 % correct on the knowledge questionnaire. Items with particularly poor responses include "Self-injury is attention seeking" while items to which the majority responded correctly include "Self-injury is a women's problem" (scored as false). This pattern of responses hints at negative attitudes and stigma in the sample rather than a lack of knowledge per se. As noted earlier, knowledge regarding NSSI may be quite idiosyncratic and based on specific personal experiences rather than epidemiological or group-level observations. Given the rapid growth in our knowledge of NSSI since the development of this measure in 2003, there would be value in developing a new measure of knowledge that more clearly taps "facts" about NSSI. This may also improve the internal consistency of the measure. Knowledge did moderate the relationship between the number of movie characters participants identified with and the frequency of NSSI. It is possible that identification with movie characters overrides a protective effect of more accurate knowledge but further work is required to ascertain the true meaning of this relationship.

Implications

Our findings offer significant implications for theory, future research and clinical work. The rich history of Social Cognitive Theory in related areas lays the ground work for exploring how the theory might be applied to understating the initiation, maintenance and cessation of NSSI, and opens the door for a range of potential mediators and moderators to be tested. For example, expectancies and self-efficacy may exert differential effects at different stages of the course of NSSI. As before, in borrowing from the research on alcohol use and dependence, a two-stage process might be at play in which expectancies are more salient predictors of NSSI early in the cycle, but with

Table 6 The relationship between number of characters participants identify with and both probability of NSSI (full sample) and frequency of NSSI (NSSI sub-sample)

	Probability of NSSI[a]				Frequency of NSSI			
	B	SE *B*	Exp (*B*)	95 % CI	*B*	SE *B*	t	95 % CI
Step one	Step $\chi^2 = 74.71$, $p < .001$				$R^2 = .06$, $F = 15.23$, $p < .001$			
Depression	.13	.04	1.14***	1.06, 1.22	.04	.01	3.61***	.02, .06
Anxiety	.01	.04	.79	.94, 1.09	.03	.01	2.02	.00, .05
Stress	.06	.05	.20	.97, .16	−.02	.02	−1.10	−.05, .01
Step two	Step $\chi^2 = .20$, $p = .65$				$\Delta R^2 = .00$, $\Delta F = .98$, $p = .32$			
No. characters[b]	−.04	.10	.96	.79, 1.16	−.04	.04	−.99	−.10, .03
Step three	Step $\chi^2 = 39.65$, $p < .001$				$\Delta R^2 = .07$, $\Delta F = 7.00$, $p < .001$			
Pos. expectancies	.01	.03	1.01	.96, 1.07	−.00	.00	−.14	−.01, .01
Neg. expectancies	−.15	.03	.86***	.81, .92	.00	.00	.83	−.01, .01
Social expectancies	.09	.08	1.09	.94, 1.27	−.00	.01	−.23	−.03, .02
Self-efficacy	−.06	.03	.95	.89, 1.00	−.03	.01	−6.733***	−.04, −.02
Knowledge	.28	.03	1.32	1.00, 1.73	.01	.02	.50	−.03, .05
Pos. attitudes	−.10	.08	.91	.77, 1.07	.00	.01	.71	−.02, .03
Neg. attitudes	−.10	.07	.91	.80, 1.03	−.01	.01	−1.21	−.03, .01
Comfort	.07	.05	1.07	.97, 1.19	.02	.01	1.92	.00, .03
Step four[c]	Step $\chi^2 = 16.86$, $p = .03$				$\Delta R^2 = .05$, $\Delta F = 5.50$, $p < .001$			
No. characters × negative expectancies	−.14	.06	.87**	.77, .97				
No. characters × social expectancies	.29	.09	1.33**	1.11, 1.60				
No. characters × knowledge					.10	.03	−4.08***	.05, .15
No. characters × positive attitudes					−.05	.01	−4.06***	−.08, −.03

** $p < .01$; *** $p < .001$

[a] History of NSSI (1 = yes, 2 = no); [b] number of characters identified with; [c] only significant interactions included in the table

repeated NSSI the behavior becomes a conditioned response to given cues (e.g., feeling distressed) with self-efficacy to resist NSSI a more salient predictor of NSSI once the behavior is entrenched (Oei and Baldwin 1994).

We have focused on NSSI-specific cognitions which might operate to explain how exposure to NSSI could increase the probability of engaging in NSSI or maintain NSSI in people already engaging in the behavior. Yet, most models of NSSI highlight the role of emotion regulation (Chapman et al. 2006; McKenzie and Gross 2014; Selby and Joiner 2009). That the positive expectancy factor of our measure of NSSI outcome expectancies contained a number of items reflecting anticipated affect regulation (e.g., I would feel better about myself) supports this. However we would argue that emotion regulation and the NSSI-specific cognitions work in concert to govern NSSI. As such we suggest future work include an examination of how cognition, experience of emotion and regulation of emotion interact to differentially predict NSSI outcomes.

In this article, we began the initial steps in developing a measure of NSSI outcome expectancies that could be used to assess anticipated outcomes of NSSI, both in people who engage in the behavior and in those who do not. Further work is required to confirm the factor structure, establish reliability and validate the scale, however this preliminary work demonstrates internally consistent scales that differentiate between people who self-injure and those who do not in a meaningful way. That the subscales differentially relate to probability of NSSI and frequency of the behavior in the sample of self-injurers offers additional construct validity, and suggests theoretical distinction between initiation and maintenance of the behavior. How expectancies and self-efficacy change over time, in line with the natural course of NSSI,[4] would be particularly informative in designing effective early intervention and treatment programs. For example, expectancy challenge interventions might be most effective in prevention and early intervention programs, while skills building and development of self-efficacy may be more relevant for more severe self-injurers.

[4] Given a large proportion of our self-injury group had not self-injured in the previous 12 months, we conducted exploratory analyses assessing differences in the cognitive variables between participants who did not self-injure, those who had not self-injured in the prior 12 months and those who had self-injured in the last year. No differences between the self-injuring groups were noted except in self-efficacy to resist NSSI.

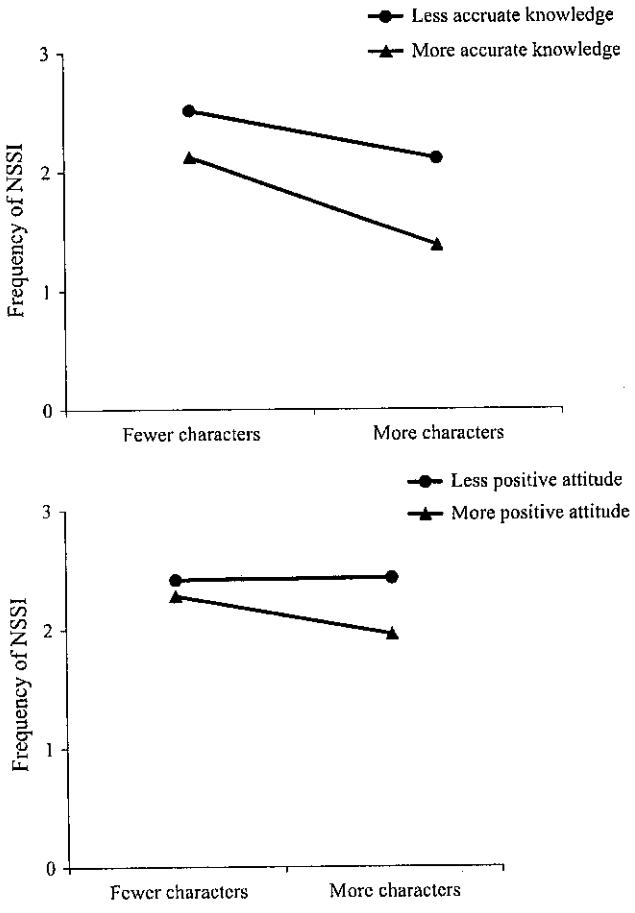


Fig. 4 Accuracy of knowledge and positive attitudes moderate the relationship between the number of movie characters participants identify with and frequency of NSSI

Limitations and Suggestions for Future Research

While the potential implications of our findings are exciting, we have only begun to embark on what is likely to be a long line of theory-driven work to better understand the initiation, maintenance and cessation of NSSI. Given that all data are self-report, from a self-selected convenience sample, our work requires replication in a larger, more representative sample, after we have further validated the new measure of NSSI outcome expectancies, and similar measures of self-efficacy. A more sophisticated approach to statistical analysis could also be adopted. Given the nested nature of the data collected (e.g., number of friends known to self-inure is dependent on whether the participant knows *any* friends who self-injure), the research questions could not be answered in a more parsimonious model. While we did adjust our alpha level to partially compensate for the number of analyses, future work would benefit from consideration of how more parsimonious models of the relationships of interest could be tested.

We proposed that the relationships between exposure to NSSI, and engagement in NSSI, night be moderated by cognitive variables. Yet, it is equally likely that these variables act to mediate this relationship. Specifically, exposure to NSSI may directly inform someone's thoughts about NSSI, which in turn would affect the probability an individual self-injures. Testing of mediation or indirect effects of exposure on engagement in NSSI is best achieved with longitudinal data that can better establish the directionality of these relationships. Experimental work in which exposure to NSSI-related stimuli is manipulated, and the subsequent cognitions assessed, is also a valuable avenue for future work. Consideration of other variables that may impact these relationships, including experience of closely related emotional disorders such as depression and anxiety, is another worthy endeavor.

Finally, it is difficult to tease apart potential contagion effects from associative relationships (in which like individuals are attracted to each other; Joiner 1999) but the application of Social Cognitive Theory offers pathways by which the potential mechanisms of action can be empirically tested. Relatedly, while we asked about identification with film characters, we did not enquire as to the nature or quality of peer relationships which may impact engagement in NSSI. We also did not define what we meant by identification with movie characters, leaving this open to interpretation. In the future, this construct could be deconstructed with clear examinations of the importance of demographic similarities, similarities in life circumstances and experiences, and whether the character is viewed in a positive or negative light.

Conclusion

In this study, we have conducted a preliminary test of whether Social Cognitive Theory can be applied to better understanding NSSI, exploring the moderating roles of core cognitive constructs in the relationship between exposure to NSSI and both history and frequency of NSSI. Overall, we propose that Social Cognitive Theory can be a useful framework that offers significant potential to explore unique predictors of onset, maintenance and cessation of NSSI, with potential for prevention, early intervention and relapse prevention programs. In particular, NSSI outcome expectancies and self-efficacy specifically related to NSSI are constructs worthy of further investigation.

Author's Contributions P.H. conceptualised the study, oversaw data collection, analysed data and prepared sections of the written manuscript. A.R. collected data, provided intellectual input to the study and prepared sections of the written manuscript. All authors have read and approved the final manuscript.

Funding This study was internally funded by The School of Psychology and Speech Pathology, Curtin University.

Conflict of interest The authors report no conflicts of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the Curtin University Human Research Ethics Committee, the National Health and Medical Research Council's National Statement on Ethical Conduct in Human Research and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

Aguinis, H., & Stone, R. E. F. (1997). Methodological artefacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology, 82*, 192–206.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. California: Sage.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179–211.

Albert, D., Chein, J., & Steinberg, L. (2013). Peer influences on adolescent decision making. *Current Directions in Psychological Science, 22*, 114–120.

Arnett, J. J. (2000). Emerging adulthood: A theory of development from the late teens through the twenties. *American Psychologist, 55*, 469–480.

Bandura, A. (1977). *Social learning theory*. New Jersey: Prentice Hall.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist, 37*, 122–147.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. NJ: Prentice Hall.

Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist, 44*, 1175–1184.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W. H. Freeman and Co.

Bentley, K. H., Barlow, D. H., & Nock, M. (2014). The four-function model of non-suicidal self-injury: Key directions for future research. *Clinical Psychological Science, 2*, 638–656.

Berger, E., Reupert, A., & Hasking, P. (2015). Pre-service and in-service teachers' knowledge, attitudes, and confidence towards self-injury among students. *Journal of Education for Teaching, 41*, 37–51.

Brown, S. A., Goldman, M. S., & Christiansen, B. A. (1985). Do alcohol expectancies mediate drinking patterns of adults? *Journal of Consulting and Clinical Psychology, 53*, 512–519.

Brown, S. A., Goldman, M. S., Inn, A., & Anderson, L. R. (1980). Expectations of reinforcement from alcohol: Their domain and relation to drinking patterns. *Journal of Consulting and Clinical Psychology, 48*, 419–426.

Chapman, A. L., Gratz, K. L., & Brown, M. Z. (2006). Solving the puzzle of deliberate self-harm: The experiential avoidance model. *Behavior Research and Therapy, 44*, 371–394.

Courtney, M. G. R. (2013). Determining the number of factors to retain in EFA: Using the SPSS R-Menu v2.0 to make more judicious estimations. *Practical Assessment, Research and Evaluation, 18*, 1–14.

Curcio, A. L., Mak, A. S., George, A. S., & George, A. M. (in press). Predictors of delinquency among adolescents and young adults: A new psychosocial control perspective. *Australian and New Zealand Journal of Criminology*.

Czyz, E. K., Bohnert, A. S., King, C. A., Price, A. M., Kleinberg, F., & Ilgen, M. A. (2014). Self-efficacy to avoid suicidal action: Factor structure and convergent validity among adults in substance use disorder treatment. *Suicide and Life-Threatening Behavior, 44*, 698–709.

Darkes, J., & Goldman, M. S. (1998). Expectancy challenge and drinking reduction: Process and structure in the alcohol expectancy network. *Experimental and Clinical Psychopharmacology, 6*, 64–76.

Deliberto, T. L., & Nock, M. K. (2008). An exploratory study of correlates, onset, and offset of non-suicidal self-injury. *Archives of Suicide Research, 12*, 219–231.

Eisenberg, D., Hunt, J., & Speer, N. (2013). Mental health in American colleges and universities: Variation across student subgroups and across campuses. *Journal of Nervous and Mental Disease, 201*, 60–67.

Greenfield, S. F., Hufford, M. R., Vagge, L. M., Muenz, L. R., Costello, M. E., & Weiss, R. D. (2000). The relationship of self-efficacy expectancies to relapse among alcohol dependent men and women: A prospective study. *Journal of Studies on Alcohol, 61*, 345–351.

Hasking, P., Andrews, T., & Martin, G. (2013). The role of exposure to self-injury among peers in predicting later self-injury. *Journal of Youth and Adolescence, 42*, 1543–1556.

Hasking, P. A., Momeni, R., Swannell, S., & Chia, S. (2008). The nature and extent of deliberate self-injury in a non-clinical sample of young adults. *Archives of Suicide Research, 12*, 208–218.

Heath, N. L., Ross, S., Toste, J. R., Charlebois, A., & Nedecheva, T. (2009). Retrospective analysis of social factors and nonsuicidal self-injury among young adults. *Canadian Journal of Behavioral Science, 41*, 180.

Heath, N. L., Toste, J. R., Somberger, M. J., & Wagner, C. (2011). Teachers' perceptions of non-suicidal self-injury in the schools. *School Mental Health, 3*, 35–43.

Joiner, T. (1999). The clustering and contagion of suicide. *Current Directions in Psychological Science, 8*, 89–92.

Jones, B. T., Corbin, W., & Fromme, K. (2001). A review of expectancy theory and alcohol consumption. *Addiction, 96*, 57–72.

Kadden, R. M., & Litt, M. D. (2011). The role of self-efficacy in the treatment of substance use disorders. *Addictive Behaviors, 36*, 1120–1126.

Klonsky, E. D., & Glenn, C. R. (2009). Assessing the functions of non-suicidal self-injury: Psychometric properties of the Inventory of Statements About Self-injury (ISAS). *Journal of Psychopathology and Behavioral Assessment, 31*, 215–219.

Klonsky, E. D., & Olino, T. M. (2008). Identifying clinically distinct subgroups of self-injurers among young adults: A latent class analysis. *Journal of Consulting and Clinical Psychology, 76*, 22.

Kokaliari, E. D. (2014). An exploratory study of the nature and extent of nonsuicidal self-injury among college women. *International Journal of Population Research*. Article ID 879269.

Lee, N. K., Greely, J., & Oei, T. P. (1999). The relationship of positive and negative alcohol expectancies to patterns of consumption of alcohol in social drinkers. *Addictive Behaviors, 24*, 359–369.

Lee, N. K., Oei, T. P., Greeley, J. D., & Baglioni, A. J, Jr. (2003). Psychometric properties of the Drinking Expectancy Questionnaire: A review of the factor structure and a proposed new scoring method. *Journal of Studies on Alcohol, 64*, 432–436.

Lewis, S. P., & Seko, Y. (2016). A double-edged sword: A review of the potential benefits and risks of online non-suicidal self-injury activities. *Journal of Clinical Psychology, 72*, 249–262.

Lovibond, S. H., & Lovibond, P. F. (1995). *Manual for the depression anxiety stress scales* (2nd ed.). Sydney: Psychology Foundation.

McKenzie, K. C., & Gross, J. J. (2014). Nonsuicidal self-injury: An emotion regulation perspective. *Psychopathology, 47*, 207–219.

Muehlenkamp, J. J., Claes, L., Havertape, L., & Plener, P. L. (2012). International prevalence of adolescent non-suicidal self-injury and deliberate self-harm. *Child abd Adolescent Psychiatry and Mental Health, 6*, 1–9.

Muehlenkamp, J. J., Claes, L., Quigley, K., Prosser, E., Claes, S., & Jans, D. (2013). Association of training on attitudes towards self-injuring clients across health professionals. *Archives of Suicide Research, 17*, 462–468.

Nixon, M. K., Cloutier, P., & Jansson, S. M. (2008). Nonsuicidal self-harm in youth: A population-based survey. *Canadian Medical Association Journal, 178*, 306–312.

Nock, M. K. (2009). Why do people hurt themselves? New insights into the nature and functions of self-injury. *Current Directions in Psychological Science, 18*, 78–83.

Nock, M. K., & Banaji, M. R. (2007). Prediction of suicide ideation and attempts among adolescents using a brief performance-based test. *Journal of Consulting and Clinical Psychology, 75*, 707–715.

Nock, M. K., Park, J. M., Finn, C. T., Deliberto, T. L., Dour, H. J., & Banaji, M. R. (2010). Measuring the "suicidal mind:" Implicit cognition predicts suicidal behavior. *Psychological Science, 21*, 511–517.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments and Computers, 32*, 396–402.

O'Connor, R. C., & Armitage, C. J. (2003). Theory of planned behavior and parasuicide: An exploratory study. *Current Psychology, 22*, 196–205.

Oei, T. P., & Baldwin, A. R. (1994). Expectancy theory: A two-process model of alcohol use and abuse. *Journal of Studies on Alcohol, 55*, 525–534.

Perkins, W. H. (2002). Social norms and the prevention of alcohol misuse in collegiate contexts. *Journal of Studies on Alcohol, Supplementary, 14*, 164–172.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *British Medical Journal, 316*, 1236–1238.

Prinstein, M. J., Heilbron, N., Guerry, J. D., Franklin, J. C., Rancourt, D., Simon, V., & Spirito, A. (2010). Peer influence and nonsuicidal self-injury: Longitudinal results in community and clinically-referred adolescent samples. *Journal of Abnormal Child Psychology, 38*, 669–682.

Radovic, S., & Hasking, P. (2013). The relationship between film portrayals of non-suicidal self-injury, attitudes, knowledge and behaviour. *Crisis, 34*, 324–334.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology, 1*, 43–46.

Schmidtke, A., & Häfner, H. (1988). The Werther effect after television films: New evidence for an old hypothesis. *Psychological Medicine, 18*, 665–676.

Selby, E. A., & Joiner, T. E. (2009). Cascades of emotion: The emergence of borderline personality disorder from emotional and behavioral dysregulation. *Review of General Psychology, 13*, 219–229.

Steinberg, L., & Monahan, K. C. (2007). Age differences in resistance to peer influence. *Developmental Psychology, 43*, 1531–1543.

Swannell, S., Martin, G., Krysinska, K., Kay, T., Olsson, K., & Win, A. (2010). Cutting on-line: Self-injury and the internet. *Advances in Mental Health, 9*, 177–189.

Swannell, S. V., Martin, G. E., Page, A., Hasking, P., & St John, N. J. (2014). Prevalence of nonsuicidal self-injury in nonclinical samples: Systematic review, meta-analysis and meta regression. *Suicide and Life Threatening Behavior, 44*, 273–303.

Tatnell, R., Kelada, L., Hasking, P., & Martin, G. (2014). Longitudinal analysis of adolescent NSSI: The role of intrapersonal and interpersonal factors. *Journal of Abnormal Child Psychology, 42*, 885–896.

Trewavas, C., Hasking, P. A., & McAllister, M. (2010). Representations of non-suicidal self-injury in motion pictures. *Archives of Suicide Research, 14*, 89–103.

Warm, A., Murray, C., & Fox, J. (2003). Why do people self-harm? *Psychological Health & Medicine, 8*, 72–79.

Whitlock, J., Muehlenkamp, J., Eckenrode, J., Purington, A., Abrams, G. B., & Barriera, P. (2013). Nonsuicidal self-injury as a gateway to suicide in young adults. *Journal of Adolescent Health, 52*, 486–492.

Wilcox, H. C., Arria, A. M., Caldeira, K. M., Vincent, K. B., Pinchevsky, G. M., & O'Grady, K. E. (2012). Longitudinal predictors of past-year non-suicidal self-injury and motives among college students. *Psychological Medicine, 1*, 1–10.

Penelope Hasking is an Associate Professor of Psychology at Curtin University. Her primary research interests concern high risk behaviour among adolescents and young adults, with a particular focus on mechanisms underlying the initiation and maintenance of nonsuicidal self-injury.

Alyssa Rose has a particular interest in the mental health of young people. She is currently a postgraduate student at Curtin University.

Empirical Research

HAMMILL INSTITUTE ON DISABILITIES

Journal of Positive Behavior Interventions
2017, Vol. 19(1) 14–25

DOI: 10.1177/1098300716632878
jpbi.sagepub.com
SAGE

Teaching Stimulus Control via Class-Wide Multiple Schedules of Reinforcement in Public Elementary School Classrooms

Jessica N. Torelli, MEd[1], Blair P. Lloyd, PhD[1], Claire A. Diekman, BA[1], and Joseph H. Wehby, PhD[1]

Abstract

In elementary school classrooms, students commonly recruit teacher attention at inappropriately high rates or at inappropriate times. Multiple schedule interventions have been used to teach stimulus control by signaling to students when reinforcement is and is not available contingent on an appropriate response. The purpose of the current study was to evaluate the effects of a class-wide multiple schedule on differentiated rates of student recruitment of teacher attention in two public elementary classrooms. General education teachers implemented the multiple schedule intervention in the context of a common instructional routine (i.e., small group rotations among reading centers). Results indicated that the multiple schedule intervention was effective at decreasing disruptive bids for attention when teacher attention was not available. Additional research on teacher implementation of class-wide multiple schedules is needed to evaluate whether this intervention may be identified as an effective supplement to Tier I classroom management strategies within multi-tiered systems of support.

Keywords

stimulus control, multiple schedule, class-wide, Tier I intervention

Teaching students appropriate ways to recruit teacher attention has been shown to increase levels of teacher praise and instructional assistance, as well as positively impact student performance in the classroom (Alber & Heward, 2000). However, once students learn how to appropriately recruit teacher attention, these behaviors can occur at inappropriately high rates or at inappropriate times (Cammilleri, Tiger, & Hanley, 2008; Sidener, Shabani, Carr, & Roland, 2006; Stokes, Fowler, & Baer, 1978). Because teacher attention is rarely continuously available to all students, it is critical not only to teach students *how* to recruit attention but *when* to recruit attention, particularly in classrooms with high student–teacher ratios.

In the applied behavior analytic literature, *multiple schedules* have been used to teach children to discriminate when adult attention is available contingent on appropriate attention-seeking behaviors. A multiple schedule is a compound schedule of reinforcement in which two or more component schedules alternate (usually according to time), and each component schedule is correlated with a different stimulus (Catania, 2013; Ferster & Skinner, 1957). Discriminative stimuli (S^D) signal the availability of reinforcement contingent on a response, whereas delta stimuli (S^Δ) signal the unavailability of reinforcement contingent on a response. Multiple schedules have been used to thin dense schedules of reinforcement following the acquisition of a target behavior. Such arrangements typically involve preserving a dense response–reinforcer contingency in one schedule component while introducing and systematically extending extinction in another schedule component. Relative to other methods of schedule thinning, multiple schedules have been found to be more effective at facilitating stimulus control while maintaining low levels of inappropriate behavior (Hanley, Iwata, & Thompson, 2001).

To address inappropriately high rates of attention recruitment, multiple schedule interventions have been applied in simulated classroom settings with preschool-age students (e.g., Tiger & Hanley, 2004, 2005; Tiger, Hanley, & Heal, 2006). Tiger and colleagues alternated between an extinction component during one stimulus (S^Δ) and a dense

[1]Vanderbilt University, Nashville, TN, USA

Corresponding Author:
Blair P. Lloyd, Department of Special Education, Peabody College, Vanderbilt University, Box 228, Nashville, TN 37203, USA.
Email: blair.lloyd@vanderbilt.edu

schedule of reinforcement (e.g., fixed ratio 1 [FR1]) during a different stimulus (S^D). Each stimulus signaled whether reinforcement (i.e., teacher attention) was available contingent on appropriate bids for attention (e.g., hand-raising or calling the teacher's name) and verbal reminders of each contingency were provided to students prior to sessions. Such multiple schedule arrangements have effectively produced discriminated responding when applied by experimenters individually and with pairs of students in small instructional rooms intended to simulate classroom settings (Tiger & Hanley, 2004, 2005; Tiger et al., 2006).

Fewer studies, however, have evaluated teacher implementation of multiple schedule interventions on a class-wide level in typical classrooms. Teaching students to discriminate when teacher attention is and is not available may be especially relevant in general education classrooms because attention is seldom continuously available for any single student. In addition, implementation of multiple schedules (i.e., pairing visual stimuli with the availability of reinforcement) requires relatively low response effort for teachers. As general education teachers are increasingly expected to assume greater responsibility in the implementation of academic and behavioral interventions (Franklin, Kim, Ryan, Kelly, & Montgomery, 2012; Horner et al., 2009; Simonsen et al., 2010), the identification of behavior support strategies that can be implemented class-wide with relative ease and efficiency is critical. Thus, multiple schedules may offer an efficient strategy to address a common problem in general education classrooms.

The identification of effective class-wide behavior support strategies is also critical for schools adopting multi-tiered systems of support (e.g., Positive Behavioral Interventions and Supports [PBIS]; Horner et al., 2009; Sugai & Horner, 2006). Class-wide strategies are designed to impact the greatest proportion of students and prevent more serious behavior problems from emerging (Horner & Sugai, 2015; Sugai et al., 2000). Multiple schedule interventions may provide an avenue to strengthen critical components of a universal classroom management plan, including maximizing structure and predictability, acknowledging appropriate behavior, and responding consistently to inappropriate behavior (e.g., systematically withholding attention; Simonsen, Fairbanks, Briesch, Myers, & Sugai, 2008). Specifically, the presentation of visual stimuli to signal the availability of teacher attention may help clarify expectations and maximize structure and predictability during instructional routines. In addition, reinforcing student bids for attention in the presence of one stimulus, and withholding attention following bids in the presence of another stimulus, would support consistent acknowledgment of attention recruitment at appropriate times without acknowledging inappropriately high rates or ill-timed bids for attention. If multiple schedule interventions can be shown to decrease disruptive bids for attention when applied class-wide by teachers, this intervention may be identified as an effective supplement to other Tier 1 classroom management strategies.

To our knowledge, only two studies have trained classroom teachers to implement multiple schedules on a class-wide basis. Cammilleri et al. (2008) trained private school teachers to implement a class-wide multiple schedule and evaluated intervention effects on student social approaches via a multiple baseline design across three private elementary school classrooms. Each classroom was comprised of 10 to 12 students ranging in age and skill level (kindergarten to sixth grade), with each student engaged in self-paced individualized curricula. Teachers conducted 5-min sessions alternating between S^D (green lei signaling teacher attention was available) and S^Δ (red lei signaling teacher attention was unavailable). Teachers implemented two to four sessions per day and described the contingency to the class before beginning each session. Results revealed that the class-wide multiple schedule produced discriminated responding such that bids for attention decreased during S^Δ, and maintained during S^D.

Vargo, Heal, Epperley, and Kooistra (2014) evaluated the effects of a class-wide multiple schedule intervention (with verbal descriptions of contingencies) on hand-raising via a multiple baseline design across three preschool classrooms. Each classroom included 18 to 19 students and the multiple schedule intervention was implemented in the context of circle time routines. Teachers alternated between signaling the availability of reinforcement for hand raises in the presence of a blue card (S^D) and the unavailability of reinforcement in the presence of a yellow card (S^Δ). Results demonstrated that hand-raising decreased in the presence of S^Δ relative to S^D across all three classrooms. In addition, responding in the S^Δ condition decreased overall during the multiple schedules plus rules component relative to baseline across all classrooms.

Because the classrooms in the study by Cammilleri and colleagues (2008) had relatively low student-to-teacher ratios, were comprised of students who ranged in age and skill level, and involved a unique instructional format (i.e., self-paced individualized instruction), it is unknown whether similar class-wide applications of multiple schedules would be effective in the context of typical public elementary classrooms. Results of the study by Vargo and colleagues (2014) suggest that multiple schedule interventions can be effectively implemented in the context of common instructional routines (i.e., circle time) in larger preschool classrooms, but these interventions have yet to be implemented in instructional routines common to elementary classrooms. In light of this small but promising literature on teacher-implemented class-wide multiple schedule interventions, additional research is needed to evaluate (a) the extent to which these procedures can be implemented with fidelity by general education teachers and (b) effects of

multiple schedules on differentiated rates of student recruitment of teacher attention in elementary classrooms. The purpose of the current study was to examine the effects of a multiple schedule procedure on stimulus control of students' recruitment of teacher attention in two general education elementary school classrooms. We addressed the following research questions:

Research Question 1: To what extent can general education teachers embed a class-wide multiple schedule intervention within an existing instructional routine with fidelity?

Research Question 2: Does teacher implementation of a class-wide multiple schedule intervention produce differentiated rates of attention recruitment for students in early elementary public school classrooms?

Method

Participants and Settings

The study was conducted in two general education classrooms in two urban, public elementary schools. The percentage of students receiving free or reduced-price lunch was 43% in one school (Classroom A) and 95% in the other school (Classroom B). Classroom A was a second-grade classroom with one general education teacher and 19 students. The general education teacher was a 34-year-old Caucasian woman with 6 years teaching experience and a master's degree with certification in general education grades K–6. The class included nine Caucasian students, eight African American students, one Latino student, and one Native American student. Three students in the class received special education services. Classroom B was a first-grade classroom with one general education teacher and 18 students. The general education teacher was a 25-year-old Hispanic woman with 3 years teaching experience and a master's degree with certification in general education grades K–6. The class included 14 African American students, three Caucasian students, and one Latino student. Three students in the class received special education services. Both classrooms contained a U-shaped small group instruction table, a teacher desk, and curricular materials. Classroom A contained 19 individual student desks, and Classroom B contained three student tables.

Each classroom used a Tier 1 classroom management strategy across all experimental conditions including baseline (i.e., Class-Wide Function-Related Intervention Teams [CW-FIT]; Kamps et al., 2011, 2015). CW-FIT is a group contingency intervention implemented on a class-wide level that focuses on decreasing disruptive behaviors by teaching socially appropriate skills and providing positive reinforcement for appropriate behaviors. In the current study, teachers had been trained to provide behavior-specific praise to teams of students who were meeting classroom expectations on average intervals of 3 to 5 min.

Both general education teachers carried out all experimental procedures during reading centers routines in the classroom. During reading centers in both classrooms, the general education teacher provided small group instruction with three to five students at a time at a U-shaped table while the remainder of the class worked at independent reading centers. In Classroom B, the teacher played music to initiate transitions between reading centers and stopped the music once the next instructional routine began; this procedure was present across all study phases. In Classroom A, a special education teacher was present during the first 20 min of the reading instruction routine and taught a small group of four to five students in the classroom. In Classroom B, the school librarian typically was present during the last 30 min of the reading centers routine and delivered instruction to a small group of five students in the classroom.

Inclusion Criteria

Teachers were recruited from schools and classrooms that had previously participated in a randomized control trial evaluating CW-FIT and were required to play CW-FIT in their classroom at the time of data collection (Kamps et al., 2011, 2015). If a teacher from the CW-FIT randomized control trial expressed interest in participating in the current study, the researcher conducted a 20-min observation to determine whether the classroom met inclusion criteria. First, to be included in the study, we required that each classroom have at least one instructional period when the teacher's attention was unavailable on a class-wide level or unavailable to a select group of students in the class (e.g., the teacher was providing small group instruction), as well as a period when the teacher's attention was available to the entire class. Second, we required a minimum of four instances of students recruiting teacher attention during the initial observation (i.e., at least one instance per 5-min interval during reading centers). Third, at least one of those instances of students recruiting teacher attention had to have occurred at an inappropriate time (i.e., when teacher attention was not available). Classrooms with no instances of inappropriate recruitment in the initial observation were excluded.

Materials

In both classrooms, a table lamp, measuring approximately 20 inches in height with a 60-W light bulb, was used as the correlated stimulus. Each lamp had been present in each classroom prior to beginning the study. We used tablets with Multi-Option Observation System for Experimental Studies (MOOSES; Tapp, Wehby, & Ellis, 1995) software to collect timed-event data on dependent and independent variables

(i.e., procedural fidelity), as well as paper-and-pencil forms to collect additional data on procedural fidelity.

Measurement System and Response Definitions

Student behavior: Recruitment of teacher attention. Data collectors scored an instance of student recruitment of teacher attention each time a student (a) raised his or her hand over the plane of the shoulder, (b) called the teacher's name, (c) handed materials to the teacher, (d) placed materials in front of the teacher, (e) made a vocal request of the teacher, (f) made a statement directed to the teacher, or (g) walked up to the teacher and stood within 3 feet of her while remaining physically oriented toward her for a minimum of 2 s (adapted from Cammilleri et al., 2008). Statements were considered directed to the teacher if the student was physically oriented toward the teacher, making eye contact with the teacher, or walking toward the teacher when the statement was made. Examples of recruitment included a student saying, "Mrs. Smith," raising his or her hand and waiting to be called on, asking the teacher to use the restroom, or walking up to the teacher and asking for help. Non-examples included calling out a question that was directed to a peer, turning in materials to a designated classroom location, responding to a teacher's statement or question, or verbal disruptions not directed to the teacher. Extended instances of recruitment (e.g., raising a hand for 2 min) were scored as a single occurrence.

During FR1 schedule components, occurrences of attention recruitment were coded across all students in the class because the FR1 schedule applied to all students in the class. During extinction components, however, extinction was in place for the subset of students in the class who were not receiving small group instruction. Thus, during S^{Δ}, instances of attention recruitment were only coded for students who were not receiving small group instruction from a teacher. Across classrooms, the number of students who received small group instruction (either from the general education teacher or other adult) ranged from three to 10.

Teacher behavior: Contingent attention. Data collectors scored contingent attention each time a teacher gestured toward or verbally acknowledged student recruitment within 5 s of the instance of recruitment (Cammilleri et al., 2008). If the student's recruitment of teacher attention took place over an extended duration of time (e.g., the student's hand remained raised or the student stood in close proximity and oriented to the teacher), contingent attention was scored if attention was provided while the student was recruiting attention or if attention was provided within 5 s of the termination of recruitment. Examples included the teacher calling the student's name, walking up to the student (within 3 feet), verbally or gesturally responding to the student's question or comment, providing a redirection, taking materials from the student's hand, patting the student on the back, or telling the student she would "be right there." Non-examples included responding to the whole class (e.g., "Remember, everyone needs to be working on his or her own assignment") or providing attention to a student who did not recruit her attention.

Data collection system. Data collectors used MOOSES (Tapp et al., 1995) to collect timed-event data on discrete instances of student recruitment (dependent variable) and teacher contingent attention (independent variable). Because session durations varied, data on student recruitment were summarized and graphed as rates (i.e., number of student recruitment responses per minute).

Procedural fidelity. From the timed-event data on student recruitment and teacher contingent attention, researchers evaluated the extent to which the teacher provided contingent attention during each component of the multiple schedule. During each FR1 and extinction session, the total frequency of contingent attention was divided by the total frequency of student recruitment and multiplied by 100% to yield a percentage of student recruitments after which attention was delivered. Thus, percentages at or near 100% during FR1 sessions and at or near 0% during extinction sessions represented correct implementation. Based on the number of extinction sessions coded with zero instances of student recruitment, we calculated mean contingent attention percentages in two ways. As summarized in Table 1, sessions in which no instances of student recruitment occurred were scored as 0% contingent attention. Additional percentages are summarized in text, however, reflecting percentages of contingent attention for the subset of sessions in which one or more instances of student recruitment occurred. Researchers also completed a checklist during each session to evaluate whether the teacher (a) explained the contingency to the class at the start of each day's reading centers routine and (b) presented the appropriate stimulus for the duration of each session. Procedural fidelity data were collected during 100% of sessions across all conditions of the study.

Implementation fidelity. The researcher collected implementation fidelity of teacher training procedures by completing a self-checklist with the following items: (a) reviewed the purpose of the study, (b) reviewed planned procedures for the FR1 schedule component, (c) reviewed planned procedures for the extinction schedule component, (d) modeled reinforcement and extinction following a student recruitment, and (e) delivered the teacher script. The researcher completed all items as planned (100%) across Classrooms A and B. Although the researcher offered each teacher an opportunity to role-play both reinforcement and extinction of student recruitment as a potential component of teacher

Table 1. Procedural Fidelity by Classroom, Experimental Condition, and Intervention Component.

	Classroom A			Classroom B		
Component	Baseline	MS (B)	MS (C)	Baseline	MS (B)	MS (C)
Correct correlated stimulus	na	93%	92%	na	96%	100%
Verbal reminder	0%	44%	43%	0%	100%	100%
Contingent attention during *light on*	55% (13–100)	4.5% (0–100)	78% (50–100)	67% (0–100)	11% (0–100)	85% (50–100)
Contingent attention during *light off*	48% (0–83)	79% (40–100)	10% (0–100)	82% (67–100)	82% (25–100)	10% (0–100)

Note. For Condition B, *light on* was correlated with extinction and *light off* was correlated with FR1. For Condition C, *light on* was correlated with FR1 and *light off* was correlated with extinction. MS = multiple schedule; FR1 = fixed ratio 1; na = not applicable.

training, neither teacher chose to role-play responses to student recruitment during the training meeting.

Inter-Observer Agreement (IOA)

Observer training included the following procedures. First, the first author met with data collectors to review the coding manual and address questions related to the coding manual and observation procedures. Next, data collectors practiced collecting data with the first author in a classroom setting until they reached a training criterion of 90% agreement for two consecutive observations.

A second observer collected data simultaneously and independently to evaluate IOA. For Classroom A, IOA data were collected during a total of 38% of sessions across S^D and S^Δ schedule components, but varied by experimental phase (33%, 36%, 69%, 59%, 17%, and 0% for Phases A, B1, C1, B2, C2, and follow-up visits, respectively). For Classroom B, IOA data were collected during a total of 49% of sessions across S^D and S^Δ schedule components but also varied by experimental phase (17%, 42%, 60%, 50%, 50%, 25%, and 33% for Phases A, C1, B1, C2, B2, withdrawal of correlated stimuli, and follow-up visits, respectively).

IOA was evaluated using a point-by-point method, in which each coded response was identified as an agreement or disagreement according to a 5-s window of agreement. Thus, for each session and dependent variable, the number of agreements (i.e., responses coded by both observers within a 5-s window) was divided by the number of agreements plus disagreements. This quotient was then multiplied by 100%. The mean percentage of agreement across conditions and observers was 94% (range = 67%–100%) for student recruitment and 97% (range = 73%–100%) for contingent attention in Classroom A. In Classroom B, the mean percentage of agreement across conditions and observers was 87% (range = 50%–100%) for student recruitment and 83% (range = 40%–100%) for contingent attention.

For additional fidelity measures (i.e., verbal reminders of contingencies and use of appropriate stimuli), agreements were scored when both observers agreed on the presence or absence of each planned teacher behavior. The number of agreements was divided by the number of agreements plus disagreements, with the quotient multiplied by 100%. We collected IOA data on these additional fidelity variables during 16% of sessions for Classroom A (during Phases B2 and C2 only) and 20% of sessions for Classroom B (during Phases B1, B2, and C2 only). Mean IOA on these fidelity measures was 100% across classrooms.

Experimental Design

We used an alternating treatments design embedded in a multi-treatment design (Birnbrauer, Peterson, & Solnick, 1974) to evaluate the effects of the multiple schedule intervention on students' recruitment of teacher attention. Specifically, FR1 and extinction schedule components were rapidly alternated within phases, and the stimuli correlated with each schedule component were reversed across phases. Reversing the stimuli correlated with each schedule component strengthened experimental control by providing additional opportunities to demonstrate effects without having to withdraw the intervention. Multiple schedule conditions B and C were counterbalanced across classrooms. Rate of student recruitment for teacher attention was the primary dependent variable used to make design-related decisions. Phase changes were implemented following (a) consistent response differentiation between FR1 (S^D) and extinction (S^Δ) conditions and (b) stable, near-zero rates of student recruitment during the extinction (S^Δ) condition.

Procedures

Baseline (A). During baseline, each teacher was instructed to follow her typical reading centers routine with the exception of turning the lamp on or off at 10-min intervals. We included these stimuli (*light on*, *light off*) during baseline to ensure that the light did not influence student responding prior to being paired with each schedule component. In both classrooms, sessions were conducted during reading centers. The general education teacher provided small group instruction to three to five students at a time at a U-shaped table while the remainder of the class rotated among independent reading centers. In addition, CW-FIT procedures

were in place during baseline and all subsequent conditions. For Classroom A, researchers collected data during twelve 10-min baseline sessions across four school days; for Classroom B, researchers collected data during eighteen 10-min baseline sessions across three school days. Observers positioned themselves in an unobtrusive location in the classroom and interacted with students to the minimum extent possible.

Teacher training. Following baseline, the researcher met with the teacher to train her on the intervention procedures. The initial training meeting lasted approximately 15 to 20 min. The researcher reviewed the purpose of the study and the procedures to implement a multiple schedule. Based on the reading center instructional rotation observed during the initial observation and baseline sessions, the researcher proposed how the multiple schedule could be embedded in the rotating centers (i.e., FR1 during transitions; extinction during small group instruction [for students who were not receiving small group instruction]). The researcher explained how to respond to student's requests for attention when the S^D was present (FR1 schedule) and when the S^Δ was present (extinction). Emphasis was placed on withholding attention following all instances of recruitment during S^Δ to ensure recruitment was not intermittently reinforced. The researcher instructed the teacher to respond to as many recruitments as possible during S^D but acknowledged that it may not be feasible to respond to each and every student recruitment. The researcher modeled reactions to instances of recruitment during reinforcement and extinction contingencies and gave the teacher the opportunity to role-play. The teacher was given a script to say aloud to the class prior to the beginning of each daily reading centers routine (see "Multiple Schedule" section). In addition, prior to each contingency reversal phase, the researcher reminded the teacher of the reversal of correlated stimuli and provided a revised script to communicate the new rules to the class.

Multiple schedule conditions. As the reading center rotations and CW-FIT procedures continued, the teacher was instructed to signal when her attention was available on an FR1 schedule (i.e., during student transitions among centers) and when her attention was not available (i.e., when she was delivering small group instruction) using the lamp. During the S^D (lamp on or off depending on condition), the teacher was instructed to do her best to respond to each instance of recruitment within 5 s. During S^Δ, the teacher was instructed to not respond to any instances of recruitment from students for whom she was not providing small group instruction. Prior to implementing the multiple schedules procedure for the class period, the teacher provided a verbal reminder of the signaled contingencies:

> For example,
>
> Remember, I'm going to use my light during centers. When my light is off, I can answer your questions and help you. When my light is on, I am working with the group at my table and I am not available to answer questions.

Sessions alternated based on these naturally occurring center rotations. On most days, teachers rotated through three (Classroom A) or four (Classroom B) small group instruction periods, which resulted in three or four sessions of each condition (i.e., FR1 and extinction) per day. Because the S^D components were programmed during center transitions, whereas S^Δ components were programmed during each reading center, S^D sessions were systematically shorter than S^Δ sessions across classrooms. For Classroom A, the mean duration of S^D sessions was 4 min (range = 1–10 min) whereas the mean duration of S^Δ sessions was 18 min (range = 7–45 min). For Classroom B, the mean duration of S^D sessions was 2 min (range = 1–5 min) whereas the mean duration of S^Δ sessions was 15 min (range = 4–24 min).

Multiple schedule (B). During multiple schedule Condition B, *light off* signaled the FR1 component during transitions among centers, and *light on* signaled the extinction component during each reading center (i.e., teacher delivery of small group instruction). This was the first multiple schedule condition introduced in Classroom A.

Multiple schedule (C). During multiple schedule Condition C, *light on* signaled the FR1 schedule component during transitions among centers and *light off* signaled the extinction component during each reading center (i.e., teacher delivery of small group instruction). This was the first multiple schedule condition introduced in Classroom B.

Withdrawal of correlated stimuli (D). In Classroom B only, researchers instructed the teacher to (a) no longer deliver the scripted announcement at the start of each daily reading centers routine and (b) stop using the lamp to signal schedule components (Phase D). This phase was conducted to evaluate whether stimulus control had transferred from the light to other naturally occurring stimuli in the environment associated with the reading centers routine. Procedures during these sessions were similar to the previous multiple schedule conditions (i.e., contingent attention provided during transitions and withheld during small group instruction), except that the lamp was not turned on and off with each change in schedule component. The lamp remained present, but the light was off for the duration of the reading centers routine. In addition, research staff instructed the teacher to try her best not to provide any verbal reminders of rules (i.e., when students could request attention) throughout the reading centers routine. This was done to evaluate the extent to which differentiated student recruitment maintained in the absence of verbal reminders or programmed correlated stimuli.

Follow-up visits. At 2 and 3 weeks following the final intervention phase, we returned to each classroom and collected data on student recruitments and teacher contingent attention during components of the reading centers routine (i.e., transitions and small group instruction). Teachers were told that data collectors would come at some point during the week but not told which day they were coming. We communicated to teachers that the study was complete and that they were free to use or not use the multiple schedule intervention according to their preference.

Social Validity

To assess teacher acceptability of intervention procedures, we asked teachers to complete a brief questionnaire designed by the first author. The questionnaire consisted of five items, and teachers rated the extent to which they agreed with each item based on a 5-point Likert-type scale (1 = *strongly disagree*, 5 = *strongly agree*). Items included the following: (a) Overall, the intervention has been effective in helping my students know when it is okay to recruit my attention; (b) The intervention was easy for me to implement; (c) I would recommend this intervention to other teachers; (d) I will continue to use this intervention after the conclusion of the study; and (e) The intervention was a helpful strategy for my classroom management. As a more objective measure of social validity, we returned to each classroom following completion of the study to determine whether each teacher chose to continue using the intervention after the study ended, as well as whether intervention effects on student behavior maintained (follow-up visits; Kennedy, 2002).

Results

Research Question 1: Teacher Implementation of Class-Wide Multiple Schedule

Classroom A. A summary of procedural fidelity data by classroom, experimental condition, and intervention component is depicted in Table 1. In Classroom A, the teacher used the correct correlated stimulus during more than 90% of sessions across multiple schedule conditions but provided verbal reminders of contingencies to a lesser degree. Specifically, verbal reminders were provided during 75% of daily reading centers routines during Phase B1, 67% of daily routines during Phase C1, 20% of daily routines during Phase B2, and 25% of daily routines during Phase C2, producing means of 44% and 43% for Conditions B and C, respectively. Mean percentages of contingent teacher attention differed between stimulus conditions correlated with each component schedule. During Condition B, in which *light on* was correlated with extinction, mean contingent attention was much lower during *light on* sessions relative to *light off* sessions (4.5% and 79%, respectively). During Condition C, in which *light on* was correlated with the FR1 schedule, mean contingent attention was much higher during *light on* sessions relative to *light off* sessions (78% and 10%, respectively). For Classroom A, one or more instances of student recruitment occurred in 100% of FR1 sessions, but in only 38% of extinction sessions. When extinction sessions with no instances of student recruitment (i.e., no opportunity for contingent attention) were removed, the mean percentage of contingent attention during extinction was 18% (range = 0%–100%). Low numbers of student recruitments per session contributed to the wide ranges in contingent attention percentages.

Classroom B. In Classroom B, the teacher used the correct correlated stimulus during more than 95% of sessions across conditions. In contrast to Classroom A, the teacher in Classroom B also provided verbal reminders of contingencies during 100% of daily reading centers routines across conditions. In fact, when we instructed the teacher in Classroom B not to refer to the contingency rules when we withdrew correlated stimuli in Phase D, she referred to contingency rules during 26% of daily routines. During Condition B, in which *light on* was correlated with extinction, mean contingent attention was much lower during *light on* sessions relative to light *light off* sessions (11% and 82%, respectively). During Condition C, in which *light on* was correlated with FR1, mean contingent attention was much higher during *light on* sessions relative to *light off* sessions (85% and 10%, respectively). For Classroom B, one or more instances of student recruitment occurred in 100% of FR1 sessions, and in 71% of extinction sessions. When extinction sessions with no instances of student recruitment were removed, the mean percentage of contingent attention during extinction was 15% (range = 0%–100%).

Research Question 2: Effects of Multiple Schedule on Student Recruitment

Classroom A. Results of the multiple schedule intervention on student recruitment in Classroom A are depicted in Figure 1. During baseline, rates of recruitment were undifferentiated during *light on* and *light off*. When the teacher initiated the multiple schedule in condition B (i.e., *light on* signaled extinction and *light off* signaled FR1), rates of recruitment decreased during *light on*, and were consistently lower relative to recruitment rates during *light off*. When the stimuli correlated with each schedule component were reversed across phases, immediate changes in level of rates of student recruitment during each stimulus were observed. In addition, beginning in Phase C1, rates of student recruitment remained at or below 0.14 during the stimulus correlated with the extinction component (S^{Δ}) across all subsequent phases.

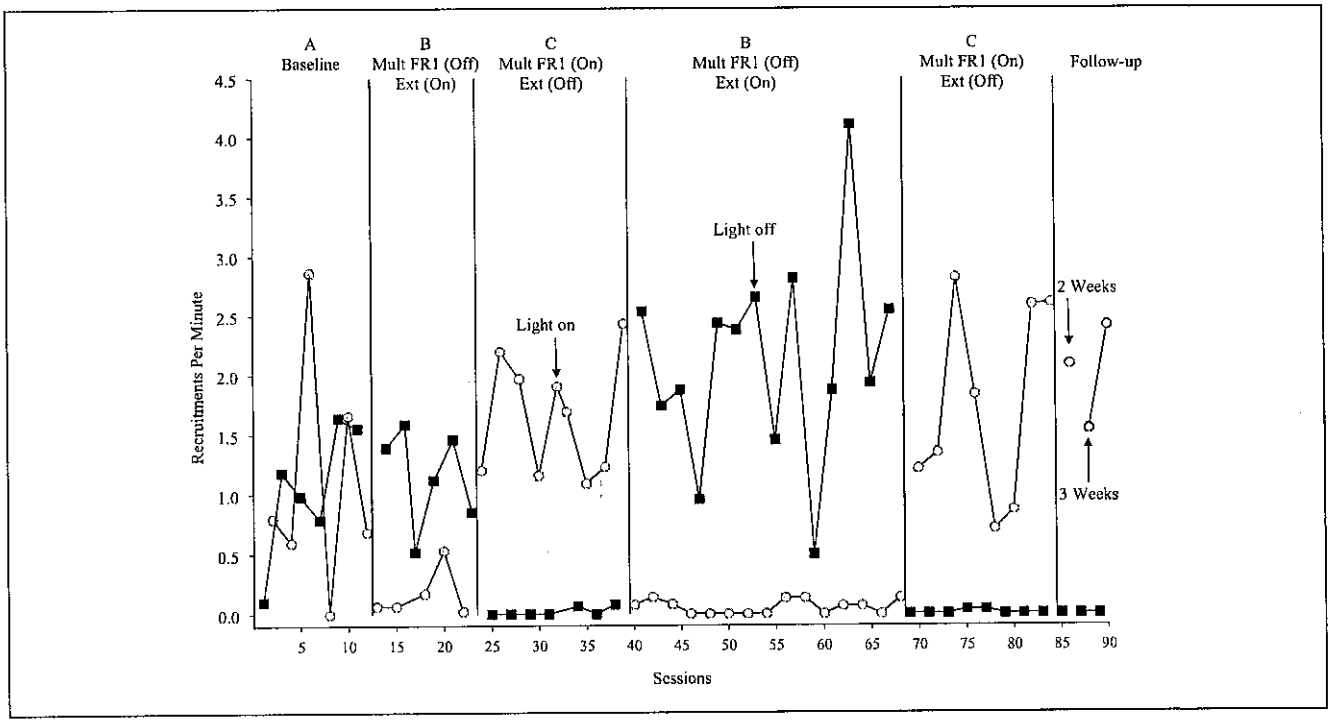


Figure 1. Rates of student recruitments of teacher attention during *light on* and *light off* stimulus conditions across study phases (Classroom A).
Note. FR1 = fixed ratio 1.

Rates of recruitment during S^D were variable across phases (range = 0.50–4.11 recruitments per min) but were consistently differentiated from recruitment rates during S^Δ across all phases. Based on the three demonstrations of effect following each reversal of correlated stimuli, we interpreted these results to indicate that the teacher-implemented multiple schedule intervention was successful in bringing student recruitment under stimulus control.

Classroom B. Results of the multiple schedule intervention on student recruitment in Classroom B are depicted in Figure 2. Similar to Classroom A, rates of recruitment were undifferentiated during *light on* and *light off* during baseline. When the teacher initiated multiple schedule Condition C (i.e., *light off* signaled extinction and *light on* signaled FR1), rates of recruitment decreased during *light off* and were less variable relative to baseline, especially as this phase continued. Unlike Classroom A, however, rates of recruitment *increased* during *light on* and were highly variable, although consistently higher relative to *light off*. When the stimuli correlated with each schedule component were reversed across phases, immediate changes in level of rates of student recruitment were observed. Rates of recruitment remained low during stimuli correlated with extinction (S^Δ) across phases, but were slightly higher relative to those observed in Classroom A (M = 0.17, range = 0–0.82). Based on the three demonstrations of effect following each reversal of correlated stimuli, we interpreted these results to indicate that the teacher-implemented multiple schedule intervention was successful in bringing student recruitment under stimulus control.

When correlated stimuli were withdrawn in Classroom B (Phase D), response differentiation persisted with lower rates of recruitment when the teacher was delivering small group instruction (extinction) relative to transitions among centers (FR1). We interpreted these data to suggest a transfer of stimulus control from the *light on*/*light off* stimuli to other natural classroom stimuli signaling the availability of teacher attention. However, as this phase progressed, a slight increasing trend was identified in the rate of student recruitment during extinction (M = 0.16, range = 0–0.40).

Descriptive baseline analysis. Because experimental data are graphed by stimulus condition (i.e., *light on* vs. *light off*) in Figures 1 and 2, the baseline data represent rates of student recruitment during 10-min sessions in which the lamp was on or off regardless of instructional context. To evaluate the practical significance of intervention effects, we also calculated baseline rates of student recruitment during small group instruction periods (i.e., reading centers) versus transitions among reading centers. Prior to introducing the multiple schedule intervention in Classroom A, the mean rate of student recruitments was 1.60 per min (range = 0.54–3.91) during transitions among centers and 1.03 per

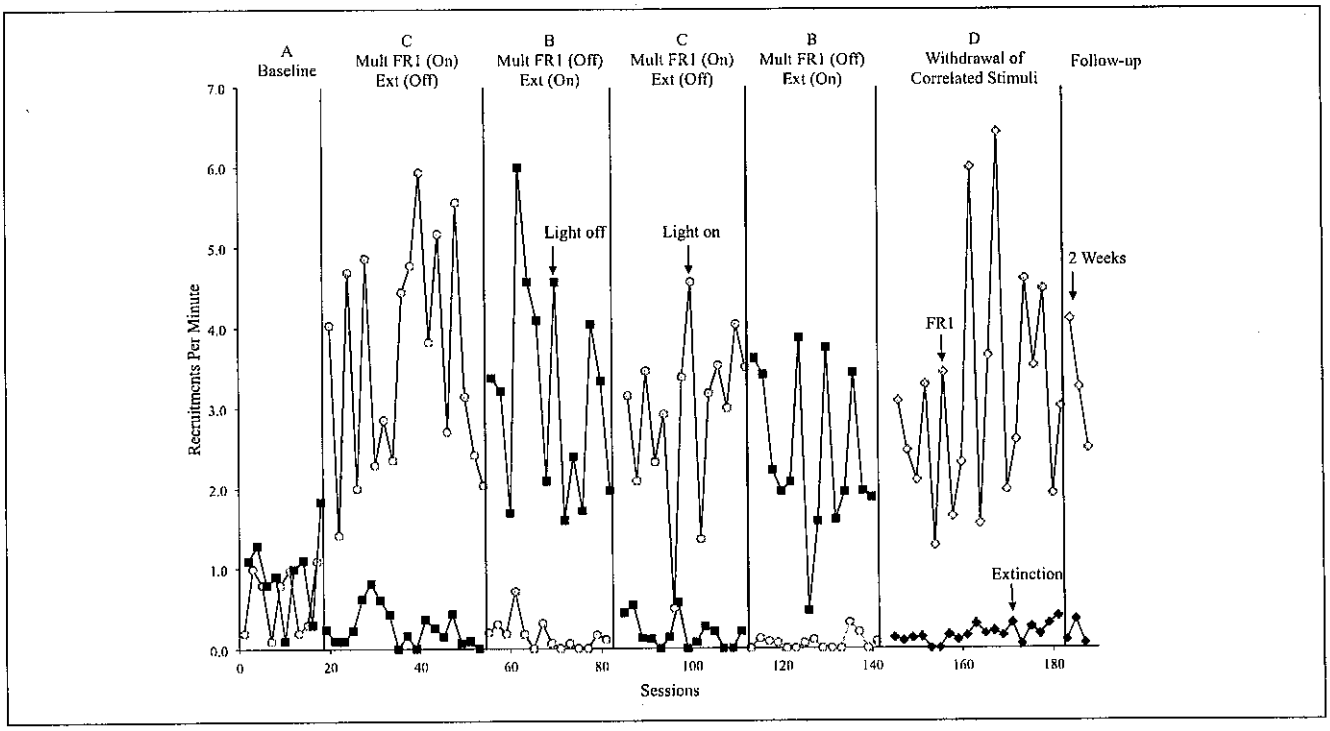


Figure 2. Rates of student recruitments of teacher attention during *light on* and *light off* stimulus conditions across study phases (Classroom B).
Note. FR1 = fixed ratio 1.

min (range = 0-4.11) during small group instruction. In Classroom B, the mean rate of student recruitments was 2.22 per min (range = 0.57–4.29) during transitions among centers and 0.60 per min (range = 0.10–1.29) during small group instruction. That is, prior to implementing the multiple schedule intervention, mean rates of student recruitment were higher during transitions relative to small group instruction. However, the addition of the multiple schedule intervention reduced rates of student recruitment during instructional periods to near-zero.

Social Validity

The multiple schedule intervention was rated by both teachers as highly acceptable. The teacher in Classroom A rated all items on the questionnaire as *strongly agree*; the teacher in Classroom B rated all items on the questionnaire as *agree* or *strongly agree*. The teacher in Classroom A continued to use the intervention at 2- and 3-week follow-up observations, and the rates of recruitment during the instructional periods remained at 0. The teacher in Classroom B did not elect to use the correlated stimuli at the follow-up observation, though she continued to withhold attention during small group instruction and respond to recruitments during transitions. Rates of recruitment during instructional periods remained at a low level (M = 0.17, range = 0.06–0.35). The teacher in Classroom B reported that she continued to use the intervention with correlated stimuli after the study concluded, with *light off* signaling the FR1 schedule, and *light on* signaling extinction.

Discussion

The purpose of the present study was to evaluate (a) the extent to which general education teachers were able to implement the multiple schedule intervention during reading centers and (b) the effects of the multiple schedule intervention on differentiation of student recruitment when teacher attention was and was not available. Results indicated that, despite varying levels of fidelity by intervention component and classroom, the multiple schedule procedure was effective in producing differentiated levels of student recruitment. That is, students learned to discriminate between periods in which teacher attention was and was not available contingent on recruitment following the addition of a stimulus (i.e., *light on*/*light off*) that was correlated with each schedule component. This study is among the first to demonstrate evidence for a class-wide application of multiple schedule interventions implemented by general education teachers in public elementary school classrooms. In addition, this study contributes an example of how multiple schedules can be embedded within a common instructional routine in elementary classrooms (i.e., small group rotations).

In the present study, differentiated rates of student recruitment per schedule component persisted across study phases, despite varying levels of fidelity by intervention component and classroom. As depicted in Table 1, there was a clear difference in fidelity between classrooms for providing verbal reminders of the contingency rules prior to each daily reading center routine. Whereas the teacher in Classroom B provided these reminders consistently across all intervention phases, the teacher in Classroom A provided these reminders less often overall and to a lesser extent across study phases. Although our design does not allow an isolation of the effects of the programmed stimuli with or without accompanying verbal reminders, such reminders are likely to facilitate stimulus discrimination when *initiating* a multiple schedule intervention (Tiger & Hanley, 2004) but may not be necessary once behavior comes under control of the correlated stimuli. Additional evidence suggesting the transfer of stimulus control was provided in Classroom B when differentiated responding continued following the withdrawal of correlated stimuli.

The multiple schedule intervention used in the present study differs from previous applications of multiple schedules in that the correlated stimuli were added to existing schedule components in a naturally occurring instructional routine. In contrast, previous applications of multiple schedule interventions implemented for individual students have commonly involved beginning with brief S^{Δ} intervals relative to S^{D} and systematically increasing the duration of the S^{Δ} component to thin the schedule of reinforcement (e.g., Hanley et al., 2001). Previous class-wide applications have involved systematically alternating schedule components in fixed intervals (e.g., 5 min FR1, 5 min Ext [Cammilleri et al., 2008]; 2–3 min FR1, 2–3 min Extinction [Vargo et al., 2014]). In the present study, the S^{Δ} and S^{D} components were embedded into existing reading center routines. Because the S^{Δ} was paired with each reading center, and the S^{D} was paired with transitions between reading centers, S^{Δ} sessions were longer than S^{D} sessions across all study phases. That is, it was not necessary to begin with brief S^{Δ} intervals and systematically increase this duration when the multiple schedule was embedded within the existing reading centers routine. However, in the current study, both participating teachers were implementing CW-FIT as a Tier 1 intervention, which included providing behavior-specific praise on average every 3 to 5 min. Whether similar intervention effects would have been observed in the absence of CW-FIT (or under conditions with lower overall rates of teacher attention) is unknown.

Limitations

Study results should be interpreted in light of the following limitations. First, though we consider embedding the multiple schedule intervention in existing instructional routines a practical advantage of the intervention, the systematic differences in durations of S^{Δ} and S^{D} sessions represents a potential threat to internal validity. In addition, because the S^{Δ} components were programmed when the teacher was delivering small group instruction, extinction was only in effect for the students who were not receiving small group instruction, resulting in systematically fewer students who had the opportunity to recruit teacher attention during S^{Δ} sessions relative to S^{D} sessions. These systematic differences in schedule component durations and number of students "present" impacted rates of recruitment per session. Rates of student recruitment were selected as the primary dependent variable to control for varying durations of sessions across schedule components. To address the difference in number of students, we completed a secondary analysis in which we divided rates of recruitment by the number of students who had the opportunity to recruit attention per session. Analyses revealed similar patterns of response differentiation across study phases for both classrooms.

Second, we did not track recruitment at the level of individual students. Instead, we collected data on recruitment rates across students to evaluate the multiple schedule intervention as a supplement to a Tier 1 strategy. However, collecting data on individual student's rates of recruitment may provide opportunities to maximize the effectiveness of the class-wide intervention. For example, anecdotally, we noted that during one session in Classroom B, a single student accounted for five of seven recruitments during the S^{Δ} component by repeatedly raising her hand. Evaluating whether specific students consistently accounted for the majority of inappropriate recruitments (i.e., recruitment during S^{Δ} components) would have allowed us to implement supplementary procedures for non-responders (e.g., pre-corrections for individual students).

Third, because each teacher implemented the multiple schedule intervention in the context of a Tier 1 classroom management strategy (i.e., CW-FIT), we do not know the extent to which our results would generalize to classrooms without Tier 1 strategies in place. A primary component of CW-FIT is a relatively dense schedule of teacher praise delivered to teams of students who are meeting expectations. Thus, it is possible that the scheduled delivery of teacher praise functioned as an abolishing operation (AO; Laraway, Snycerski, Michael, & Poling, 2003) for recruiting teacher attention during the reading centers routine. However, the relatively high rates of attention recruitment that maintained during S^{D} components across intervention phases suggest that the value of teacher attention provided to individual students remained high. Finally, although effects of the multiple schedule intervention are consistent with the hypothesis that teacher attention was a reinforcer for student recruitment, we did not conduct reinforcer assessments or confirm response–reinforcer relations prior to implementing the multiple schedule.

Future Research

Additional research on class-wide applications of multiple schedules is needed to evaluate whether this intervention may be identified as an effective supplement to Tier 1 strategies. In addition, several questions that were not addressed in the current study could be addressed in future research. First, we observed consistent intervention effects despite variability in fidelity across components and classrooms. Future studies might systematically manipulate fidelity levels of critical components to evaluate the levels of fidelity necessary to impact student behavior. Second, although there was some evidence suggesting a transfer of stimulus control from verbal rules to correlated stimuli (Classroom A) and from correlated stimuli to natural stimuli associated with the reading centers routine (Classroom B), future research may evaluate systematic strategies for transferring stimulus control (e.g., systematically fading out correlated stimuli following response differentiation). Third, because the purpose of the multiple schedule was to decrease disruptive bids for attention thought to interrupt instruction, incorporating measures of academic engagement in future studies may provide additional insight into whether multiple schedules ultimately impact the time spent academically engaged. Fourth, although we included follow-up visits as a measure of social validity, future studies might evaluate the extent to which intervention effects maintain over longer periods of time and generalize when correlated stimuli are introduced during other instructional routines. Finally, data collection on response differentiation at the level of individual students would allow an identification of students who are not responding to the intervention. For these students, it would be important to identify effective secondary or tertiary strategies to increase the likelihood of treatment response.

Conclusion

Although multiple schedules have been applied to individuals in clinical settings to facilitate stimulus control of a newly acquired response, this study is among the first to demonstrate evidence for a class-wide application of multiple schedules implemented by general education teachers in public elementary school classrooms. With continued research, class-wide multiple schedules may be identified as an evidence-based strategy that can be embedded successfully within common instructional routines to enhance effects of Tier 1 interventions.

Authors' Note

This research was part of the first and third authors' theses submitted in partial completion of the MEd degree at Vanderbilt University.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported, in part, by the Institute of Education Sciences, Department of Education (R324A120344).

References

Alber, S. R., & Heward, W. L. (2000). Teaching students to recruit positive attention: A review and recommendations. *Journal of Behavioral Education, 10*, 177–204.

Birnbrauer, J. S., Peterson, C. R., & Solnick, J. V. (1974). Design and interpretation of studies of single subjects. *American Journal of Mental Deficiency, 79*, 191–203.

Cammilleri, A. P., Tiger, J. H., & Hanley, G. P. (2008). Developing stimulus control of young children's requests to teachers: Classwide applications of multiple schedules. *Journal of Applied Behavior Analysis, 41*, 299–303.

Catania, A. C. (2013). *Learning* (5th ed.). Cornwall-on-Hudson, NY: Sloan Publishing.

Ferster, C. B., & Skinner, B. F. (1957). *Schedules of reinforcement*. Acton, MA: Copley.

Franklin, C. G. S., Kim, J. S., Ryan, T. N., Kelly, M. S., & Montgomery, K. L. (2012). Teacher involvement in school mental health interventions: A systematic review. *Children and Youth Services Review, 34*, 973–982.

Hanley, G. P., Iwata, B. A., & Thompson, R. H. (2001). Reinforcement schedule thinning following treatment with functional communication training. *Journal of Applied Behavior Analysis, 34*, 17–38.

Horner, R. H., & Sugai, G. (2015). School-wide PBIS: An example of applied behavior analysis implemented at a scale of social importance. *Behavior Analysis in Practice, 8*, 80–85.

Horner, R. H., Sugai, G., Smolkowski, K., Eber, L., Nakasato, J., Todd, A. W., & Esperanza, J. (2009). A randomized, wait-list controlled effectiveness trial assessing school-wide positive behavior support in elementary schools. *Journal of Positive Behavior Interventions, 11*, 133–144.

Kamps, D., Wills, H., Dawson-Bannister, H., Heitzman-Powell, L., Kottwitz, E., Hansen, B., & Fleming, K. (2015). Class-wide function-related intervention teams "CW-FIT" efficacy trial outcomes. *Journal of Positive Behavior Interventions, 17*, 134–145.

Kamps, D., Wills, H. P., Heitzman-Powell, L., Laylin, J., Szoke, C., Petrillo, T., & Culey, A. (2011). Class-wide function-related intervention teams: Effects of group contingency programs in urban classrooms. *Journal of Positive Behavior Interventions, 13*, 154–167.

Kennedy, C. H. (2002). The maintenance of behavior change as an indicator of social validity. *Behavior Modification, 26*, 594–604.

Laraway, S., Snycerski, S., Michael, J., & Poling, A. (2003). Motivating operations and terms to describe them: Some further refinements. *Journal of Applied Behavior Analysis, 36*, 407–414.

Sidener, T. M., Shabani, D. B., Carr, J. E., & Roland, J. P. (2006). An evaluation of strategies to maintain mands at practical levels. *Research in Developmental Disabilities, 27*, 632–644.

Simonsen, B., Fairbanks, S., Briesch, A., Myers, D., & Sugai, G. (2008). Evidence-based practices in classroom management:

Considerations for research to practice. *Education and Treatment of Children, 31*, 351–380.

Simonsen, B., Shaw, S. F., Faggella-Luby, M., Sugai, G., Coyne, M. D., Rhein, B., . . . Alfano, M. (2010). A school-wide model for service delivery: Redefining special educators as interventionists. *Remedial and Special Education, 31*, 17–23.

Stokes, T. F., Fowler, S. A., & Baer, D. M. (1978). Training preschool children to recruit natural communities of reinforcement. *Journal of Applied Behavior Analysis, 11*, 285–303.

Sugai, G., & Horner, R. H. (2006). A promising approach for expanding and sustaining the implementation of school-wide positive behavior support. *School Psychology Review, 35*, 245–259.

Sugai, G., Horner, R. H., Dunlap, G., Hieneman, M., Lewis, T. J., Nelson, C. M., . . .Ruef, M. (2000). Applying positive behavior support and functional behavioral assessment in schools. *Journal of Positive Behavior Interventions, 2*, 131–143.

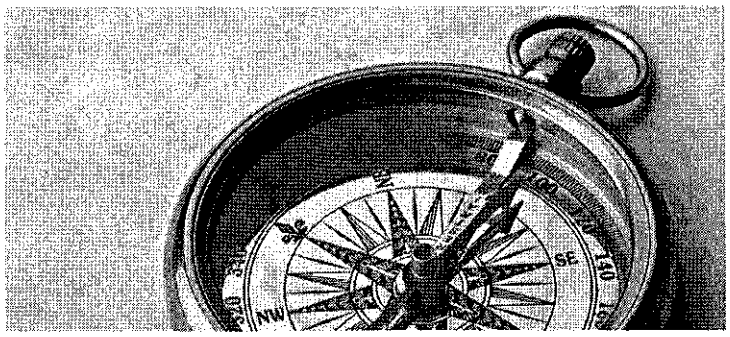
Tapp, J., Wehby, J., & Ellis, D. (1995). A multiple option observation system for experimental studies: MOOSES. *Behavior Research Methods, Instruments, & Computers, 27*, 25–31.

Tiger, J. H., & Hanley, G. P. (2004). Developing stimulus control of preschooler mands: An analysis of schedule-correlated and contingency-specifying stimuli. *Journal of Applied Behavior Analysis, 37*, 517–521.

Tiger, J. H., & Hanley, G. P. (2005). An example of discovery research involving the transfer of stimulus control. *Journal of Applied Behavior Analysis, 38*, 499–509.

Tiger, J. H., Hanley, G. P., & Heal, N. A. (2006). The effectiveness of and preschoolers' preferences for variations of multiple-schedule arrangements. *Journal of Applied Behavior Analysis, 39*, 475–488.

Vargo, K. K., Heal, N. A., Epperley, K., & Kooistra, E. (2014). The effects of a multiple schedule plus rules on hand raising during circle time in preschool classrooms. *Journal of Behavioral Education, 23*, 326–343.



Ethical Principles of Psychologists and Code of Conduct

Including 2010 and 2016 Amendments

Additional Resources

- 2018 APA Ethics Committee Rules and Procedures (PDF, 197KB)
- Revision of Ethics Code Standard 3.04 (Avoiding Harm)
- APA Ethical Principles of Psychologists and Code of Conduct (2017) (PDF, 272KB)
- 2016 APA Ethics Committee Rules and Procedures
- Revision of Ethical Standard 3.04 of the "Ethical Principles of Psychologists and Code of Conduct" (2002, as Amended 2010) (PDF, 26KB)
- 2010 Amendments to the 2002 "Ethical Principles of Psychologists and Code of Conduct" (PDF, 39KB)
- Compare the 1992 and 2002 Ethics Codes

ASSIGNMENT

Learning and cognition is a broad discipline. For this Discipline-Based Literature Review, you will research at least two peer-reviewed articles published within the last ten years to support your analysis of each of the following topics:

- Operant conditioning
- Classical conditioning
- Behaviorism
- Social learning theory

When conducting your research, consider the issues that arise within these different disciplines. In your analysis of each topic, explain the theoretical perspectives and empirical research that are pertinent to the field of learning and cognition. Within this analysis, apply ethical principles to the evaluation of each theoretical approach and consider issues that arise in research related to that area. Support your statements with evidence from your selected articles.

The Discipline-Based Literature Review

- Must be six to eight double-spaced pages in length, not including the title or references pages, and formatted according to APA style (Links to an external site.) as outlined in the Ashford Writing Center.
- Must include a title page with the following:
  - Title of paper
  - Student's name
  - Course name and number
  - Instructor's name
  - Date submitted
- Must begin with an introductory paragraph.
- Must address the topic of the paper, including each construct, with critical thought.
- Must end with a conclusion paragraph.
- Must use at least eight scholarly sources (two per topic) in addition to the required resources.
  - The Scholarly, Peer Reviewed, and Other Credible Sources (Links to an external site.) table offers additional guidance on appropriate source types. If you have questions about whether a specific source is appropriate for this assignment, please contact your instructor. Your instructor has the final say about the appropriateness of a specific source for a particular assignment.
- Must document all sources in APA style as outlined in Citing Within Your Paper (Links to an external site.).
- Must include a separate references page that is formatted according to APA style as outlined in Formatting Your References List (Links to an external site.).

Operant conditioning.

operant conditioning

Authors: Rholetter, Wylene, MEd

Source: Salem Press Encyclopedia, 2019. 2p.

Document Type: Article

Subject Terms: Operant conditioning

Abstract: Operant conditioning, a term coined by B. F. Skinner, American psychologist and radical behaviorist, is the idea that behavior is the learned result of consequences. Skinner, who introduced the concept in his 1938 book The Behavior of Organisms: An Experimental Analysis, theorized that operant conditioning in the form of reinforcements and punishments leads to an association between a behavior and its consequence. Positive reinforcement increases a desirable behavior by following it with a favorable stimulus. Negative reinforcement increases a desirable behavior by removing an unfavorable stimulus after the behavior is performed. Both positive and negative reinforcement seek to increase a desirable behavior. Punishment, like reinforcement, also has positive and negative varieties. Positive punishment is adding an unfavorable stimulus in an effort to eradicate an undesirable behavior. Negative punishment is removing an unpleasant stimulus in order to decrease undesirable behavior. Both positive and negative punishment seek to decrease an undesirable behavior.

Full Text Word Count: 959

Accession Number: **89677602**

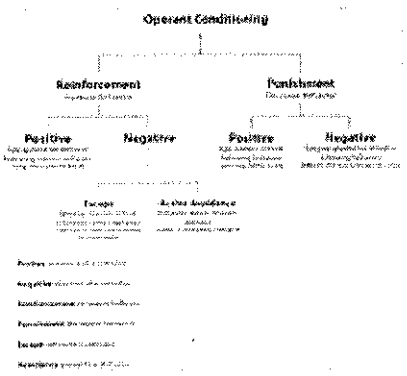
Database: Research Starters

Operant conditioning

Listen American Accent

Operant conditioning, a term coined by B. F. Skinner, American psychologist and radical behaviorist, is the idea that behavior is the learned result of consequences. Skinner, who introduced the concept in his 1938 book *The Behavior of Organisms: An Experimental Analysis,* theorized that operant conditioning in the form of reinforcements and punishments leads to an association between a behavior and its consequence. Positive reinforcement increases a desirable behavior by following it with a favorable stimulus. Negative reinforcement increases a desirable behavior by removing an unfavorable stimulus after the behavior is performed. Both positive and negative reinforcement seek to increase a desirable behavior. Punishment, like reinforcement, also has positive and negative varieties. Positive punishment

is adding an unfavorable stimulus in an effort to eradicate an undesirable behavior. Negative punishment is removing an unpleasant stimulus in order to decrease undesirable behavior. Both positive and negative punishment seek to decrease an undesirable behavior.



The hierarchy of operant conditioning. By Studentne [GFDL (http://www.gnu.org/copyleft/fdl.htn or CC-BY-SA-3.0 (http://creativecommons.org/licens sa/3.0)], via Wikimedia Commons

Overview

Skinner designed an operant conditioning chamber, which came to be known as the Skinner box, to test his theory of operant conditioning on animals. The Skinner box prevented human interruption of the experimental session and allowed the experimenter to study the behavior of an animal as a continuous process. The box includes at least one lever or key that the animal can manipulate to release food, water, or some other reward or to avoid punishment such as an electric shock. Skinner's experiments with rats and pigeons showed that the animals first hit the lever and released food accidentally; after a few accidental releases, the reinforcement of manipulating the lever ensured that the behavior would be repeated. Skinner believed that operant conditioning could be used in similar ways with human beings.

Modifying behavior through operant conditioning has been used in the treatment of phobias, obsessive-compulsive disorders, substance-abuse problems, and some sexual disorders, but the impact of Skinner's theories about operant conditioning has proved to be immense, reaching far beyond the field of psychology. Zoos and other animal facilities routinely use food as a positive reinforcement to train animals to move within enclosed areas and to increase safety during veterinary examinations. With human subjects, operant conditioning has been used to control absenteeism in the workplace (such as when employers offer staff members with no absences a chance to win cash rewards), to increase sales (coupons), and to manage agitation in older adults with dementia. Perhaps no field has been more influenced by operant conditioning than education. Skinner's assertion that positive reinforcement is more effective than punishment at changing and establishing desirable behavior led to the discrediting of punitive punishment in schools and the common application of timeouts (negative reinforcement) and a token economy (i.e., rewarding good behavior with gold stars that can be accumulated for prizes) instead.

Critics of operant conditioning have been vehement in pointing out its detriments. As early as 1959, American linguist and cognitive scientist Noam Chomsky argued that what worked in Skinner's laboratory could be applied to complex human behavior only in a superficial way. In 1960 progressive educator A. S. Neil insisted that rewarding good behavior taught that the behavior was not worth doing for reasons other than the reward. Other critics were even more severe, charging that operant conditioning was dangerous and inhumane. Gradually, the influence of Skinner's ideas declined, and by the twenty-first century, some declared that operant conditioning had become peripheral in psychology

and related fields. However, in 2002 a list of ninety-nine top psychologists was published in the *Review of General Psychology* and B. F. Skinner topped the list.

Bibliography

Bunzli, Samantha, David Gillham, and Adrian Esterman. "Physiotherapy-Provided Operant Conditioning in the Management of Low Back Pain Disability: A Systematic Review." *Physiotherapy Research International* 16.1 (2011): 4–19. *Academic Search Premier.* Web. 7 Aug. 2013.

Davey, Graham, and Chris Cullen. *Human Operant Conditioning and Behavior Modification*. New York: Wiley, 1988. Print.

Dayan, Peter. "Instrumental Vigour in Punishment and Reward." *European Journal of Neuroscience* 35.7 (2012): 1152–68. *Academic Search Premier.* Web. 7 Aug. 2013.

Dayan, Peter, et al. "Disentangling the Roles of Approach, Activation and Valence in Instrumental and Pavlovian Responding." *PLoS Computational Biology* 7.4 (2011): 1–28. *Academic Search Premier.* Web. 7 Aug. 2013.

Edwards, Darren J. *Integrating Behavioural and Cognitive Psychology: A Modern Categorization Theoretical Approach.* Hauppage: Nova, 2015. Print.

Fonseca, Amilcar Rodrigues, Maria Cristina Zago Castelli, and Emileane Costa Assis de Oliveira. "Effects of Chronic Mild Stress on Operant Discrimination Learning." *Behavior Analysis: Research and Practice* 15.1 (2015): 20–27. Print.

Iversen, Iver H. "Skinner's Early Research: From Reflexology to Operant Conditioning." *American Psychologist* 47.11 (1992): 1318–28. *PsycINFO.* Web. 24 July 2013.

Miller, Harold L., Jr., and E. Benjamin H. Heuston. "Recent Trends in Operant Conditioning." *21st Century Psychology: A Reference Handbook.* Eds. Stephen F. Davis et al. Los Angeles: Sage, 2008, 340–50. Print.

Murphy, Eric S., and Frances K. McSweeney. *The Wiley-Blackwell Handbook of Operant and Classical Conditioning.* Hoboken: Wiley, 2014. Print.

Parrish, Margaret. "Behaviorism." *Social Work Perspectives on Human Behavior.* Maidenhead: Open UP, 2010, 98–109. Print.

Rapanelli, Maximiliano, Luciana Romina Frick, and Bonifacio Silvano Zanutto. "Learning an Operant Conditioning Task Differentially Induces Gliogenesis in the Medial Prefrontal Cortex and Neurogenesis in the Hippocampus." *PLoS ONE* 6.2 (2011): 1–12. *Academic Search Premier.* Web. 7 Aug. 2013.

Reynolds, George Stanley. *A Primer of Operant Conditioning*. Rev. ed. Glenview: Scott, 1975. Print.

Staddon, J. E. R., and D. T. Cerutti. "Operant Conditioning." *Annual Review of Psychology* 54 (2003): 115–44. *PsycINFO.* Web. 24 July 2013.

Person-Centered & Experiential Psychotherapies
Vol. 10, No. 3, September 2011, 198–209

Routledge
Taylor & Francis Group

Operant Conditioning

Bridging the positions of Rogers and Skinner: The role of nonlinear dynamic systems

Lisbeth Sommerbeck*

Clinical psychologist, Bornholm Psychiatric Center, Denmark

(*Received 11 December 2010; final version received 21 March 2011*)

This paper aims to bridge the gap between the positions of Carl Rogers and B. F. Skinner by demonstrating that they are not as antithetical as they are ordinarily assumed to be, and as the two men themselves tended to think. The author uses the concept of nonlinear dynamic systems as the principal concept in her effort to bridge the gap between the work of Rogers and Skinner and consequently introduces the concept of nonpredictivity to replace the concept of nondirectivity as being a more correct and basic characterization of classical client-centered therapy. The concept of nonpredictivity allows client-centered therapy to be framed, in an overarching and general way, by B. F. Skinner's principle of operant conditioning without doing violence to Rogers' concern about client autonomy. More specifically, the paper argues that Rogers' theory of therapy is not antithetical to Skinner's theory of personality and that client-centered therapy therefore can be practiced seamlessly by therapists who, like the author, believe that operant conditioning is a better explanatory concept for human behavior than the actualizing tendency.

Keywords: Rogers and Skinner; nonlinear dynamic systems; nonpredictive; nondirective

Brückenschlag zwischen den Positionen von Rogers und Skinner: die Rolle nicht-linearer dynamischer Systeme

Dieser Artikel will die Kluft zwischen den Positionen von Rogers und Skinner schliessen. Diese sind nicht so anti-thetisch, wie man normalerweise annimmt und wie die beiden selbst meinten. Ein Konzept nicht-linearer dynamischer Systeme soll die Kluft zwischen Rogers' und Skinners' Werk schliessen. Das Konzept der Nicht-Vorhersagbarkeit soll das Konzept der Nicht-Direktivität ersetzen; ersteres ist eine korrektere und grundlegendere Charakterisierung der klassischen Klientzentrierten Therapie. Das Konzept der Nicht-Vorhersagbarkeit ermöglicht es, Klientzentrierte Therapie in umfassender und genereller Weise in Skinners Prinzip des Operanten Konditionierens einzubetten, ohne dass man Rogers' Sorge um die Autonomie der Klienten Gewalt antut. Rogers' Therapietheorie ist keine Antithese zu Skinners Theorie der Persönlichkeit und Klientzentrierte Therapie kann daher nahtlos von Therapeuten praktiziert werden, die wie die Autorin überzeugt sind, dass Operantes Konditionieren ein besseres Erklärungskonzept für menschliches Verhalten ist als die Aktualisierungstendenz.

*Email: lisbethsommerbeck@privat.dk

ISSN 1477-9757 print/ISSN 1752-9182 online

DOI: 10.1080/14779757.2011.599516
http://www.informaworld.com

Estableciendo un puente entre las posturas de Rogers y Skinner: El rol de los sistemas dinámicos ni lineales

Este escrito tiene como fin establecer un puente entre las posiciones de Carl Rogers y B. F. Skinner al demostrar que no son tan antagónicas como normalmente se las considera, y como ambos autores tendían a considerarlas. La autora utiliza el concepto de sistemas dinámicos no lineares como el concepto principal en su esfuerzo para salvar la distancia entre la obra de Rogers y la de Skinner, y por ello introduce el concepto de no predictibilidad para reemplazar el concepto de no directividad, considerándolo una caracterización más exacta del la terapia centrada en el cliente clásica. El concepto de no predictibilidad permite que ubiquemos la terapia centrada en el cliente en un marco más abarcativo y general, por el principio de Skinner de condicionamiento operativo sin violentar la preocupación de Rogers acerca de la autonomía del cliente. Más específicamente, este escrito argumenta que la teoría de Rogers acerca de la terapia no es contraria a la teoría de la personalidad de Skinner y que por lo tanto la terapia centrada en el cliente puede ser aplicada por terapeutas, como la autora, creen que el condicionamiento operante es un concepto que explica el comportamiento humano mejor que la tendencia actualizante.

L'élaboration d'un pont entre les positions de Rogers et de Skinner: Le rôle de systèmes non-linéaires et dynamiques

Cet article vise à créer un pont entre les positions de Carl Rogers et B.F. Skinner en démontrant qu'elles ne sont pas aussi antithétiques qu'on le suppose d'ordinaire ou que les deux hommes avaient tendance à les penser. L'auteure utilise le concept de systèmes non-linéaires et dynamiques comme concept principal dans son effort de créer un pont entre le travail de Rogers et de Skinner. Il s'ensuit qu'elle propose le concept de non-prédictibilité pour remplacer le concept de non-directivité en tant que caractérisation plus juste et plus fondamentale de la thérapie centrée sur le client classique. Le concept de non-predictibilité permet à la thérapie centrée sur le client d'être encadrée, d'une manière globale et générale, par le principe de conditionnement opérant de Skinner sans faire violence à la préoccupation de Rogers concernant l'autonomie du client. Plus spécifiquement, l'article soutient que la théorie de la thérapie de Rogers n'est pas antithétique par rapport à la théorie de la personnalité de Skinner et que la thérapie centrée sur le client peut donc être pratiquée sans heurt par des thérapeutes qui, comme l'auteure, croient que le conditionnement opérant est un concept qui explique mieux le comportement humain que la tendance actualisante.

Fazendo a ponte entre as posições de Rogers e de Skinner: O papel dos sistemas dinâmicos não lineares

Este artigo pretende estabelecer uma ponte entre as posições de Carl Rogers e de B.F. Skinner, ao demonstrar que os dois autores não são antitéticos como habitualmente se julga e como eles próprios tendiam a ver-se um ao outro. A autora recorre ao conceito de sistemas dinâmicos não lineares como sendo central no estabelecimento dessa ponte entre os trabalhos de Rogers e de Skinner e, consequentemente, introduz o conceito de não previsibilidade para substituir a não directividade, considerando o primeiro como permitindo uma caracterização mais correcta e básica da terapia centrada no cliente clássica. O conceito de não previsibilidade permite que se enquadre a terapia centrada no cliente, de forma mais abrangente e geral, através do princípio skinneriano de condicionamento operante, sem violentar a preocupação de Rogers com a autonomia do cliente. Mais concretamente, este artigo pretende que a teoria da terapia de Rogers não é antagónica com a teoria da personalidade de Skinner e que, consequentemente, a terapia centrada no cliente pode ser colocada em prática, sem problemas, por terapeutas que, como o autor, acreditam que o condicionamento operante é um

conceito mais adequado para o comportamento humano do que a tendência actualizante.

ロジャーズとスキナーの立場の橋渡し：非線形力動的システムの役割について

本論文の目的は、カール・ロジャーズとB.F.スキナーの立場が「対立的」ではないことを示し、彼らの間にある相違を埋めることである。ロジャーズとスキナーの業績の間にある相違を埋めるための重要な概念として、非線形力動的システムを用いた。古典的な来談者中心療法におけるより適切で基本的な特徴として、"非指示性"の代わりに"非予測性"を提案した。そして、B.F.スキナーのオペラント条件付けの原理による"非予測性"の概念により、来談者中心療法は、クライエントの自主性に関するロジャーズの懸念を害することなく、より包括的で普遍的に形作られるとし、ロジャーズのセラピーに関する理論がスキナーのパーソナリティ理論と正反対のものではないことを示した。そして、実現傾向よりもオペラント条件付けが人間の行動の説明概念として優れていると信じている治療者であったとしても、来談者中心療が全ての学派にわけ隔てなく実践されうると論じた。

En brobygger mellem Rogers og Skinner: De ikke-lineære dynamiske systemers rolle

Denne artiikel har til formål at bygge bro mellem Carl Rogers' og B. F. Skinners positioner ved at demonstrere at de ikke er så anti-tetiske som almindeligt antaget og som de to mænd selv anså dem for at være. Forfatteren bruger begrebet "ikke-lineære dynamiske systemer" som sit vigtigste begreb i forsøget på at bygge bro mellem Rogers' og Skinners teorier, og som en konsekvens heraf introducerer hun begrebet "ikke-forudsigende" til erstatning for begrebet "ikke-direktiv," idet hun anser førstnævnte for at være en mere basal og korrekt karakterisering af klassisk klient-centreret terapi. Begrebet "ikke-forudsigende" tillader klient-centreret terapi at blive kontekstualiseret, på overordnet og generel vis, af B. F. skinners princip om operant betingning uden at gøre vold på Rogers' omsorg for klientens autonomi. Mere specifikt argumenterer artiklen for at Rogers' teori om terapi ikke er antitetisk til Skinners personlighedsteori og for at klient-centreret terapi derfor kan praktiseres gnidningsfrit af terapeuter der, som forfatteren, tror at operant betingning er et bedre begreb til forklaring af menneskers adfærd end den aktualiserende tendens.

Introduction

Writing this paper has been close to the heart of the author who finds so much of value in Rogers' as well as in Skinner's works that it has become increasingly inconceivable to comprehend their views as being excluding of or in contradiction with each other. Rogers and Skinner were, after all, among the giants of 20th century psychology. It therefore seems most likely that much of value can be found in both men's work.

If Rogers and Skinner could have had a third dialogue, today, informed by late 20th century advances in work with nonlinear dynamic systems (see below), they

might have ended up on the same page and there might not have been the split between the behaviorist orientation and the humanist orientation in psychology that we see today. They agreed (Rogers & Skinner, 1956, 1962/1989) that human behavior is determined, although Rogers had some reservations, because he found that Skinner attached too little importance to the inner, subjective experiences of the individual, particularly the experience of being free and having choices. Rogers did not clarify, though, what kind of importance should be attached to these inner, subjective experiences, apart from stating in various ways that they were important because they felt important and meaningful to the individual. He spoke of a "paradoxical" relationship between scientific determinism and inner, subjective experiences. He said (1962/1989, p. 132):

> I am in thorough agreement with Dr. Skinner that, viewed from the external, scientific, objective perspective, man is determined by genetic and cultural influences. I have also said that in an entirely different dimension, such things as freedom and choice are extremely real ... I see it as being similar to the situation in physics where you can prove that the wave theory of light is supported by evidence, as is the corpuscular theory, though the two of them appear to be contradictory. They are not, at the present state of knowledge, reconcilable, but one would be narrowing his perception of physics to deny one or the other. It is in this same sense that I regard both of these dimensions as real, although they exist in paradoxical relationship. [Physicists, today, would not speak of a paradoxical relationship, but of a complementary relationship between the wave theory and the corpuscular theory of light. (This author's comment)]

However, talking about a paradoxical relationship between scientific determinism and the experience of freedom doesn't really explain anything and, significantly, Rogers did not state that he attached causative agency to inner experiences, or to the experience of being free to make choices, that is, to the experience of having a free will. On the contrary, he said (Rogers, 1962/1989, p. 99): "Behaviors that were formerly dealt with as though they were caused by little homunculi within the individual, or various internal causes, are now seen to have other types of causation."

Rogers' reservation to Skinner's views seems first and foremost to originate in Rogers' feeling of no less than revulsion (1962, p. 126) at the thought that human behavior could be controlled *ad modum* Walden Two (Skinner, 1948). Skinner took the position that as much as he acknowledged the existence of inner, subjective experiences, including the experience of choosing freely, he regarded them as more or less covert or private behaviors, and it was the environmental consequences of a person's behavior, not any kind of inner experiences, that should be studied, because those consequences were the causes (in the form of rewards or punishments) of the behavior. By engineering the environmental consequences of a person's behavior, it was, according to Skinner, possible to control this person's behavior and that prospect was as positive from Skinner's point of view as it was negative from Rogers' point of view.

Basically, however, they were in agreement about the assumption that determinism implies predictability. Today, they would have been informed about Chaos Theory and nonlinear dynamic systems, that is, they would have known that determinism does not necessarily imply predictability. In the light of this knowledge, they might have appreciated that their respective positions were, indeed, compatible. Skinner might have understood that in spite of his being right about the principle of operant conditioning and determinism for living organisms, in general, he could not hope to be able to control the rich and complex behavior of any human individual, in

particular. A Walden Two could never be realized. Rogers, on his side, might have understood that his agreeing with Skinner about determinism did not exclude the soundness and relevance of his nonpredictive therapeutic approach with any individual client.

The impossibility of predictive control of complex sets of behavior

A basic assumption

It is a basic assumption of this paper that the trajectory of the process of the individual human organism's interaction with its environment from conception to death is best modeled by a *nonlinear dynamic system* unlike, for example, the trajectory of the planets around the sun, which is, as physics has long since demonstrated, best modeled by a linear dynamic system.

Nonlinear dynamic systems[1]

Classical Newtonian science is almost exclusively concerned with linear dynamic systems. In such systems, cause–effect relationships can be clearly and exhaustively defined. This allows for very precise predictions of the behavior of a linear dynamic system. With measurements of a few parameters, the trajectory of a planet, for example, can be very precisely predicted. This is so even if no set of measurements can be 100% precise. Physicists and mathematicians say about linear dynamic systems that small differences in initial conditions remain small differences in the future behavior of the system. For all practical purposes, therefore, small differences in initial conditions can be disregarded. Stated otherwise, one can make one's measurements as precise as one wishes one's predictions to be precise. This is the essentially characterizing feature of linear dynamic systems.

Such is not the case with nonlinear dynamic systems. They are said to be sensitively dependent on initial conditions. This means that small differences in initial conditions can develop into big differences in the future behavior of the systems. *The butterfly effect* is the ubiquitously used term for this sensitive dependence on initial conditions. To illustrate what nonlinear dynamic systems are about, it is said that a butterfly flapping its wings in Beijing on Sunday can be decisive for the weather in New York next Saturday. This means that the development of the weather around the globe is a nonlinear dynamic system, which is sensitively dependent on initial conditions, and it is the reason meteorologists cannot predict the weather with better than chance precision for much more than four or five days ahead.[2]

To the surprise of most people, mathematics, that is, the area that is normally thought of as the most predictable of all, is full of nonlinear dynamic systems. Many mathematical equations show sensitive dependence on initial conditions. If one iterates (develops) such equations on a calculator that delivers results with nine decimal points' precision and on a calculator that delivers results with eight decimal points' precision one will find that the iterations on the two calculators sooner or later start to vary widely and seemingly chaotically, as a result of the tiny, one millionth part, difference in initial conditions. *Chaos Theory* is the field of mathematics that is concerned with nonlinear dynamic systems (see, e.g., Gleick, 1987; Lorenz, 1993; Peitgen, Jürgens, & Saupe, 1992). The name is somewhat unfortunate, since there is really nothing "chaotic" about these systems. They are

fully determined systems. The meaning of the term *chaotic* is, more precisely, *unpredictable*. Nonlinear dynamic systems are as determined as linear dynamic systems, but they are as unpredictable as linear dynamic systems are predictable.

Here is another example of a nonlinear dynamic system:

Imagine that you let go of a small bullet at the edge of a large funnel with a hole in the bottom, just a tiny bit larger in diameter than the bullet. You can, of course, on the basis of the law of gravity, predict that it will end up disappearing down the hole in the bottom, but there is no way you can predict its trajectory toward that disappearance. That trajectory is sensitively dependent on the tiniest differences in the way you hold the bullet towards the edge of the funnel at the start, on the tiniest differences in where you hold it, on the tiniest differences in the way you let go of it, and on the tiniest details of irregularities on the side of the funnel. Any of these tiny differences can have a huge effect on the bullet's trajectory. At the same time, these differences are so tiny that there is no way to measure them, just as there is no way to measure their combined differential influence on the bullet's behavior in a way that is sufficiently precise to make it possible to predict the trajectory of the bullet towards its disappearance. One can only make the prediction (obvious to everyone) that the bullet's overall trajectory will be downwards, that it will probably not disappear right away, but instead hit the edge of the hole in the bottom, and be spiraled up again, one time or several times, but still closer to the bottom and still more slowly as it loses energy, until it finally disappears.

At this point, the reader is welcome to associate to the human organism's trajectory through life, since my point with this rather detailed explanation of nonlinear dynamic systems is, as already stated, to postulate that the behavior of the human organism, that is, its continuously changing complex sets of interactions with a continuously changing environment, from conception to death, is a nonlinear dynamic system, and, probably, *the* nonlinear dynamic system, par excellence. The human organism in its environment is, more than anything, like the bullet in the funnel that was described above: Its trajectory through life is as sensitively dependent on initial conditions as the bullet's trajectory on the side of the funnel.

Determined and unpredictable

The consequence of this assumption is that the interactional processes of human beings with their environment are determined *and* unpredictable. It is very important to grasp the idea that something can be both determined and unpredictable. It is important, because our language and thinking is so suffused with the linearity of classical science that we automatically tend to think of determined systems as being synonymous with predictable systems. It is therefore no wonder that Rogers and Skinner did not question the basic, mistaken, assumption that they had in common: that determinism invariably implies predictability. It was only with the advent of the modern computer that mathematicians and physicists could investigate nonlinear dynamic systems thoroughly, and knowledge about nonlinear dynamic systems did not become common knowledge until the two last decades of the past century. At the time of the dialogues between Rogers and Skinner, thinking in terms of predictable, linear dynamic systems was still the way of thinking that dominated the sciences.

The individual level and the general level

Rogers and Skinner, though, were not altogether wrong about their assumption that determinism implies predictability. They were only wrong on the level of the individual human organism's interaction with the environment, and they were not very careful about distinguishing between the individual and the general level in their dialogues. Their failure to do so may have contributed to their thinking of themselves as being in disagreement. On the general level, that is, on the level of people, at large, or of sufficiently large groups of people, or on the level of "the average," they were right. And it may be this level Rogers had in mind when he spoke of the "scientific perspective" in the quotation above. On this level Skinner's laws of conditioning can be applied to predict which environments will produce a certain kind of behavior among the majority of people exposed to this environment. And on this level Rogers and Skinner were actually in rather close agreement about the general kind of environment that would produce the most constructive behaviors in the majority of people or, vice versa, they were in agreement about the kind of environment that hampered the actualization of these behaviors, that is, they were in agreement about a general theory of disturbed psychological development or, in Skinner's terminology, disturbed behavior. For Rogers, psychological disturbance came about as a result of the individual having been excessively exposed to conditional regard. For Skinner, disturbed behavior came about as a result of the individual having been excessively exposed to punishing consequences.

Skinner's theory operates with four reinforcers, two positive (increasing the likelihood of recurrence of a certain set of behaviors) and two negative (decreasing the likelihood of recurrence of a certain set of behaviors). The two positive reinforcers are (1) finding rewarding consequences, and (2) escaping from punishing consequences; and the two negative reinforcers are (3) finding punishing consequences, and (4) finding that rewarding consequences have disappeared (O'Donohue & Ferguson, 2001, p. 92).

For Rogers conditional positive regard is actually implicitly punishing of behaviors that are not regarded positively: "I love you when you ...," implies that "I do not love you when ... " and withdrawal of love can be as painful to humans as electric shock to rats.

Thus, it is hard to see any essential difference between "conditional regard" and "punishing consequences."

Rogers, however, employed the concept "unconditional positive regard" that has no counterpart in Skinner's work, and this concept of Rogers', although he didn't explicitly say so anywhere, embraces the unpredictability of human behavior that went beyond Skinner's theory. (See further, below.) Rogers, more than Skinner, had a "sense" of nonlinear dynamics, before knowledge of these dynamics became accessible.

In any case, they were in agreement about the essential point: that disturbed psychological development or disturbed behavior is the result of unfortunate environments, that is, it is conditioned and thus controlled by the environment, even if it is impossible to tell, on the individual level, what the specific conditioning contingencies or reinforcement relationships of any given set of disturbed behaviors might have been and may be. The logical conclusion is that they were also in agreement that any behavior, not only disturbed behavior, is controlled by the environment. The

evidence is Rogers' statement, in the debate with Skinner, that "I am in thorough agreement with Dr. Skinner that, viewed from the external, scientific, objective perspective, man is determined by genetic and cultural influences" (1962/1989, p. 132; see the full quotation above). In another context Rogers wrote:

> The question is often raised: But what about the therapist's attitude toward his client's asocial or antisocial behavior? Is he to accept this without evaluation? Sometimes this question is answered by saying that the effective therapist prizes the person, but not necessarily his behavior. Yet it is doubtful if this is an adequate or true answer. To be sure, the therapist may feel that a particular behavior is socially unacceptable or socially bad, something he could not approve of in himself, and a way of behaving which is inimical to the welfare of the social group. But the effective therapist may feel acceptant of this behavior in his client, not as desirable behavior, but as a *natural consequence* of the circumstances, experiences, and feelings of this client. Thus the therapist's acceptance may be based upon this kind of feeling: "If I had had the same background, the same circumstances, the same experiences, it would be inevitable in me, as it is in this client, that I would act in this fashion." (Rogers, Gendlin, Keisler, & Truax, 1967, p. 103)

Here Rogers is fully on the same wavelength as Skinner.

Rogers' dilemma

As we have seen, Rogers, however, had problems with this and he also often spoke about the self-governing, internally controlled individual. For example, in his definition of the actualizing tendency (Rogers, 1959, p. 196) he spoke of development "toward autonomy and away from heteronomy, or control by external forces." It seems to be this apparent contradiction in Rogers' view that he tried to resolve with reference to the "paradoxical" relationship between the individual experience of freedom and the notion that human behavior is determined.

Rogers also expressed his efforts to integrate the idea of individual freedom on the one hand and determinism on the other in the following quote:

> The fully functioning person ... not only experiences, but utilizes, the most absolute freedom when he spontaneously, freely, and voluntarily chooses and wills that which is also absolutely determined. (Rogers, 1961, p. 193)

It should be evident from these quotations that Rogers struggled hard with the free will/determinism issue without resolving it. The author wonders why Rogers seemingly did not pay attention to the everyday occurrence that subjective experience and objective knowledge can be held simultaneously in mind: We can enjoy the beautiful sunset while knowing full well that the sun is not at all "setting"; instead the earth is revolving around it. Likewise, we can hold in mind, simultaneously, the enjoyment of freedom while knowing full well that the experience of freedom is a subjective experience that is as determined as everything else and certainly not an experience enjoyed by all.

Skinner's mistake

It is undoubtedly the impossibility of making sufficiently precise predictions about the individual conditioning contingencies of a given client's behavior that is at the heart of the problems of behavior analytic therapy, that is, the therapeutic

application of Skinner's laws of conditioning on the level of individual clients. The engineering of specific reinforcement strategies to change a given "target behavior" demands isolation of the target behavior and prediction about the reinforcement strategies to be applied to change the target behavior, and this is, more often than not, impossible with the complex kind of behaviors and problems ordinary clients present to ordinary practicing therapists. The rare client may present with sufficiently simple and well-delineated kinds of symptom/problem/target behavior for effective reinforcement strategies to be successfully predicted and applied. In these cases, it is possible to approach the client's "presenting problem" to a linear dynamic system and to work with it accordingly. To remain with the bullet in the funnel analogy, this would roughly correspond to a whole in the bottom that was sufficiently large to make it obvious that the bullet would disappear directly into it independently of any differential influences on its way to the bottom. As early as the middle of the last century numerous studies, referred to by Patterson (1963), showed that carefully chosen, well-delineated and obvious "bits" of behavior can be changed by predictive operant conditioning. Ordinarily, though, this kind of approximation to a linear dynamic system is not possible in therapy without disregarding essential features of the relatively complex problems that clients typically see therapists about. These problems are part of a unique nonlinear dynamic system that is as unpredictable as it is determined. With ordinary clients, the diameter of the hole in the bottom is not much larger than the diameter of the bullet, the side of the funnel is very irregular and there are an infinite number of variations in the ways different bullets were started on their unique trajectories toward the bottom.

Therefore, Chaos Theory and the basic unpredictability of nonlinear dynamic systems falsify Skinner's assumption of the predictability of the behavior of the individual human organism. Rogers' opposition to this point of Skinner's ideas seems to have been, primarily, ethically motivated, but later developments in mathematics and physics came to his support: Rogers' idea that a therapist should not try to control the behavior of any individual client gains momentum by the idea that the therapist, quite simply, *cannot* control the behavior of any given client.

A third dialogue

Thus, if Rogers and Skinner could have had a third dialogue, today, Rogers, in the light of the basic unpredictability of nonlinear dynamic systems, might not have been as uncomfortable about Skinner's deterministic position as he was in 1957 and 1962. Rogers might more unequivocally have agreed with Skinner that the human organism is determined and Skinner might have agreed that Rogers was right that the complex sets of interaction of a client with his or her environment should not be exposed to efforts of direction or control in therapy; not, primarily, because such efforts were ethically wrong, but because such efforts were, more likely than not, doomed to failure, since they would imply therapist capacity to make appropriate predictions about the client at odds with the trajectory of client changes being best modeled by a nonlinear dynamic system.

Classical client-centered therapy: Nonpredictive rather than nondirective

Classical client-centered therapy is regarded by most practitioners as being utterly nondirective (see, e.g., Bozarth, 1998, p. 57; Grant, 1990; Levitt & Brodley, 2005). It

is the therapy of the "early Rogers" or "Rogers-1" (Frankel & Sommerbeck, 2005, 2007) that Rogers described in *Counseling and Psychotherapy* (1942), where he distinguished between "directive" approaches and his own "nondirective" approach (1942, pp. 115–128). He articulated the philosophical foundation of this therapy in more detail in *Client-Centered Therapy* (1951). Nevertheless, most would probably agree that classical-client centered therapists do direct, in the sense of influence, their clients, just as clients influence their therapists, and just as we all influence each other all the time. Skinner would say that control or conditioning is mutual; it goes both ways. Therefore, the essence of nondirectivity in classical client-centered therapy is, more precisely, that the therapist does not *systematically* try to influence or direct the client. The therapist has no preconceived, that is, *predictive* idea, based on diagnostic assessment, a particular theory of personality, or any other perspective of their own, of how and in which direction they wish to influence the client, and the therapist entertains no such ideas at any moment during the course of therapy. In short, the therapist makes no effort to predict what will be helpful to a given client, but engages instead in continuous empathic understanding of the client's perspective.

This does not mean that expert predictions do not exist. However, expert predictions are made on the basis of experts' knowledge of what characterizes some average of a group of people or some sufficiently large group. Such predictions can be made with satisfying precision when the average process and the actual process are virtually the same, which is the case with linear dynamic systems like, for example, the trajectory of the planets around the sun. But for nonlinear dynamic systems, where small differences in initial conditions can develop into very large differences at a later point in time, it is impossible to know how widely the known average and the actual system of interest at this precise moment in time differ from each other, except in very crude and banal ways. Psychologists and psychiatrists, for example, are experts at how a sufficiently large group of people with a certain diagnosis responds to this or that intervention, or how the "average" person with this diagnosis responds, but have no way of knowing how close any given client with the given diagnosis is to this average (Sommerbeck, 2004, pp. 295–297). Therefore:

> Any intervention with the process of another, which is made with the intention of being helpful, as will most often be the case in therapy, is made on the basis of a prediction that this intervention will actually, more likely than not, be helpful for the other. Otherwise, the intervention would not be made. In making such predictions, the other is treated as a linear dynamic system, as "an average person" or as "an average client" on the basis of theory, research, or the therapist's professional or personal experience. The other is not treated as this particular, unique, nonlinear dynamic system, John, who is different from any average, or as this other, different and equally unique, nonlinear dynamic system, Jenny, whose future interactions with her environment it is impossible to make predictions about. (Sommerbeck, 2004, p. 296)

To this author, Rogers' fourth condition of the six necessary and sufficient conditions for therapeutic change (Rogers, 1959, p. 213), unconditional positive regard for the client, is, first and foremost, respect for the uniqueness of each client, or respect for the client being a nonlinear dynamic system that it is impossible to make anything but banal and crude, that is, more often than not therapeutically irrelevant, predictions about. It is also respect for the importance of clients' subjective *experience* of freedom, while knowing that objectively the client is as

determined as the therapist or anything/anyone else, and that the therapy, as a whole, is but another determining or conditioning influence, albeit a nonpredicted influence, in the life of the client and in the life of the therapist. The only prediction made by the classical client-centered therapist is that the client will, more likely than not, be better off at the conclusion of the therapy than at the start of it, but what this means, in any detail, is left blowing in the wind.

A personal note

This paper represents the way the author, for herself, has tried to resolve Rogers' "paradox" regarding freedom and determinism, and the way she can effortlessly and satisfyingly contain the wisdom of Rogers' client-centered therapy and the wisdom of Skinner's principle of operant conditioning at the same time. It is the way for the author to be dedicated to Rogers' theory of therapy, while being critical of Rogers' theory of personality, particularly the concept of the actualizing tendency (Frankel, Sommerbeck, & Rachlin, 2010) and instead embracing Skinner's behaviorist theory of personality.

Conclusion

Hopefully, this paper has demonstrated that if Rogers and Skinner could have had a third dialogue, today, informed by knowledge of nonlinear dynamic systems, they could have bridged the apparent gap between their respective positions. They could have agreed that, rather ironically, the very point they used to agree about, that determinism implies predictability, was the very point they were both wrong about. They would have known that most aspects of a human being's trajectory through life are a nonlinear dynamic system, determined and unpredictable, and in the light of this knowledge Rogers could have agreed unequivocally with Skinner about operant conditioning and determinism without fearing the consequences of his agreement for the relevance of his therapeutic approach. And in this same light, Skinner could have agreed about the wisdom of nonpredictive client-centered therapy for the ordinary complexity of client problems that demand unconditional respect for client uniqueness rather than respect for client freedom in some nonscientific "dimension."

A corollary of this conclusion is the dispensability of a belief in the existence of the actualizing tendency as the foundation for practicing client-centered therapy. A belief in, and respect for client uniqueness suffices.

Notes

1. This explanation of nonlinear dynamic systems is a slightly revised and shortened excerpt of Sommerbeck (2004). The author is grateful to the editors of the PCEP journal for their permission to insert this excerpt in the present article.
2. It would be more correct to say that *the trajectory of* the weather is *best modeled by* a nonlinear dynamic system, since the concept of a nonlinear dynamic system is a mathematical concept that consists of mathematical variables and their interrelationships, which the weather, of course, does not consist of. For easier readability, though, the author has chosen the less mathematically correct formulation that this or that is a nonlinear dynamic system rather than the more circumstantial, but also more correct, formulation that the trajectory of this or that is best modeled by a nonlinear dynamic system.

References

Bozarth, J. (1998). *Person-centered therapy: A revolutionary paradigm.* Ross-on-Wye, UK: PCCS Books.

Frankel, M., & Sommerbeck, L. (2005). Two Rogers and congruence: The emergence of therapist-centred therapy and the demise of client-centred therapy. In B. Levitt (Ed.), *Embracing non-directivity* (pp. 40–61). Ross-on-Wye, UK: PCCS Books.

Frankel, M., & Sommerbeck, L. (2007). Two Rogers: Congruence and the change from client-centered therapy to we-centered therapy. *Person-Centered & Experiential Psychotherapies, 4*, 286–295.

Frankel, M., Sommerbeck, L., & Rachlin, H. (2010). Rogers' concept of the actualizing tendency in relation to Darwinian theory. *Person-Centered & Experiential Psychotherapies, 9*, 69–80.

Gleick, J. (1987). *Chaos: Making a new science.* New York: Viking Penguin.

Grant, B. (1990). Principled and instrumental non-directiveness in person-centered and client-centered therapy. *Person-Centered Review, 5*, 77–88.

Levitt, B., & Brodley, B.T. (2005). "It enlightens everything you do": Observing non-directivity in a client-centred therapy demonstration session. In B. Levitt (Ed.), *Embracing non-directivity* (pp. 96–112). Ross-on-Wye, UK: PCCS Books.

Lorenz, E. (1993). *The essence of chaos.* Seattle, WA: University of Washington Press.

O'Donohue, W., & Ferguson, K. (2001). *The psychology of B. F. Skinner.* London: Sage.

Patterson, C. (1963). Control, conditioning and counselling. *Personnel and Guidance Journal, 41*, 680–686.

Peitgen, H., Jürgens, H., & Saupe, D. (1992). *Chaos and fractals.* New York: Springer-Verlag.

Rogers, C.R. (1942). *Counseling and psychotherapy.* Boston: Houghton Mifflin.

Rogers, C.R. (1951). *Client-centered therapy.* Boston: Houghton Mifflin.

Rogers, C.R. (1959). A theory of therapy, personality, and interpersonal relationships as developed in the client-centered framework. In S. Koch (Ed.), *Psychology: A study of a science. Vol. 3: Formulations of the person and the social context.* New York: McGraw-Hill.

Rogers, C.R. (1961). *On becoming a person.* London: Constable.

Rogers, C.R., & Skinner, B. (1956). Some issues concerning the control of human behavior. *Science, 124*(3231), 1057–1066.

Rogers, C.R., & Skinner, B.F. (1989). Dialogue. In H. Kirschenbaum & V. L. Henderson (Eds.), *Carl Rogers: Dialogues* (pp. 79–152). Boston: Houghton Mifflin. (Original work published 1962).

Rogers, C.R., Gendlin, E., Kiesler, D., & Truax, C. (1967). *The therapeutic relationship with schizophrenics.* Madison, WI: The University of Wisconsin Press.

Skinner, B.F. (1948). *Walden Two.* Indianapolis, IN: Hackett Publishing Company.

Sommerbeck, L. (2004). Non-linear dynamic systems and the non-directive attitude in client-centered therapy. *Person-Centered & Experiential Psychotherapies, 3*, 291–299.

Eur Arch Psychiatry Clin Neurosci (2016) 266:291–305
DOI 10.1007/s00406-015-0593-1

ORIGINAL PAPER

Classical conditioning in borderline personality disorder: an fMRI study

Annegret Krause-Utz[1,2,3] · Jana Keibel-Mauchnik[1] · Ulrich Ebner-Priemer[1,4] · Martin Bohus[1] · Christian Schmahl[1]

Classical Conditioning

Received: 16 May 2014 / Accepted: 15 March 2015 / Published online: 27 March 2015

Abstract Previous research suggests disturbed emotional learning and memory in borderline personality disorder (BPD). Studies investigating the neural correlates of aversive differential delay conditioning in BPD are currently lacking. We aimed to investigate acquisition, within-session extinction, between-session extinction recall, and reacquisition. We expected increased activation in the insula, amygdala, and anterior cingulate, and decreased prefrontal activation in BPD patients. During functional magnetic resonance imaging, 27 medication-free female BPD patients and 26 female healthy controls (HC) performed a differential delay aversive conditioning paradigm. An electric shock served as unconditioned stimulus, two neutral pictures as conditioned stimuli (CS+/CS−). Dependent variables were blood-oxygen-level-dependent response, skin conductance response (SCR), and subjective ratings (valence, arousal). No significant between-group differences in brain activation were found [all p(FDR) > 0.05]. Within-group comparisons for $CS+_{unpaired}$ > CS− revealed increased insula activity in BPD patients but not in HC during early acquisition; during late acquisition, both groups recruited fronto-parietal areas [p(FDR) < 0.05]. During extinction, BPD patients rated both CS+ and CS− as significantly more arousing and aversive than HC and activated the amygdala in response to CS+. In contrast, HC showed increased prefrontal activity in response to CS+ > CS during extinction. During extinction recall, there was a trend for stronger SCR to CS+ > CS in BPD patients. Amygdala habituation to $CS+_{paired}$ (CS+ in temporal contingency with the aversive event) during acquisition was found in HC but not in patients. Our findings suggest altered temporal response patterns in terms of increased vigilance already during early acquisition and delayed extinction processes in individuals with BPD.

Keywords Borderline personality disorder · Classical conditioning · Prefrontal cortex · Amygdala · Insula

Annegret Krause-Utz and Jana Keibel-Mauchnik have contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00406-015-0593-1) contains supplementary material, which is available to authorized users.

✉ Christian Schmahl
christian.schmahl@zi-mannheim.de

1 Department of Psychosomatic Medicine and Psychotherapy, Medical Faculty Mannheim, Central Institute of Mental Health Mannheim, Heidelberg University, J 5, 68159 Mannheim, Germany

2 Institute of Psychology, Leiden University, Leiden, The Netherlands

3 Leiden Institute for Brain and Cognition, Leiden, The Netherlands

4 Department of Sport and Sport Science and House of Competence, KIT - Karlsruhe Institute of Technology, Karlsruhe, Germany

Introduction

Conditioning is a basic associative learning process that plays an important role in the development and maintenance of psychiatric disorders as well as in behavior therapy, especially exposure therapy.

Numerous studies have applied experimental fear conditioning paradigms to examine the (neurobiological) mechanisms underlying associative emotional learning [1–31]. In differential delay conditioning paradigms, one of the two initially neutral stimuli (CS+) is temporally paired with an aversive unconditioned stimulus (US), while the other is not (CS−). Many experimental settings do not only

involve *acquisition* (learning) and within-session extinction but also between-session extinction (*extinction recall*) and *reacquisition* or *reinstatement*, i.e., a renewed exposure to the original CS–US contingency.

There is convergent evidence from both animal research and human studies that the amygdala, insula, and medial prefrontal cortex (mPFC) including the anterior cingulate cortex (ACC) play central roles in the acquisition of fear responses [1, 4, 5, 9, 15]. Several studies in healthy humans further reported differential responses (to CS+ > CS−) in the inferior frontal gyrus, rostral, and caudal orbitofrontal cortex (OFC), thalamus, hippocampus, cerebellum, striatum (including putamen), as well as in sensory cortices [6, 10, 16, 18, 20, 22, 24–28]. The mPFC is assumed to play a central role in the integration of sensory and affective stimuli and in the coordination of memory storage both during the acquisition and the extinction of emotional (fear) memories [11, 14, 28, 31]. Extinction is thought to involve activity-dependent potentiation of synaptic transmission in the mPFC resulting in an inhibition of amygdala-dependent responses [11, 14]. Animals with lesions of the mPFC are unable to extinguish conditioned responses [4, 15]. In healthy humans, increased activation in the mPFC [20, 24, 26], OFC [17], amygdala (e.g., [7, 18, 20]), and insula [2, 17] could be observed during extinction (for a review see [28]). For the amygdala, a rapid habituation, i.e., decrease in reactivity to unpleasant stimuli was observed over the course of the acquisition phase [7, 8].

Within a 'brain network of emotion regulation,' the amygdala and insula have been critically implicated in the processing of negative emotions, while the ACC is thought to play a critical role in both the generation and regulation of emotions [9, 32]; prefrontal areas including the ventrolateral and dorsolateral PFC (vlPFC, dlPFC), OFC, and dorsomedial PFC are thought to be involved in the down-regulation of affective arousal [32].

In anxiety disorders, acquired fear responses to an initially neutral stimulus (e.g., place, object, person) previously paired with an US appears to be not successfully extinguished [33–35]. Patients with posttraumatic stress disorder (PTSD), for example, showed amygdala hyperreactivity and increased skin conductance responses to the CS+ during extinction recall [34]. Deficits in associative emotional learning are assumed to underlie clinical features such as a persistent fear and heightened affective arousal, even in contexts where no actual threat is present.

To our knowledge, no study so far has investigated the neural correlates of aversive differential delay conditioning in patients with borderline personality disorder (BPD). BPD is a severe psychiatric disorder with high rates of trauma and pronounced difficulties in emotion regulation including affective hyperarousal even in normative neutral situations [36, 37]. Experimental research in patients with BPD suggests altered learning and memory in aversive emotional contexts [38, 39]. On the neural level, there is growing evidence for both structural and functional alterations in fronto-limbic brain regions which play a major role in affecting regulation and learning processes (e.g., amygdala, hippocampus, anterior cingulate) including amygdala hyperactivity and diminished recruitment of the prefrontal areas during emotional challenge (for reviews see [40, 41]).

In a previous psychophysiological study [42], we investigated skin conductance responses (SCR) as well as arousal and valence ratings during an aversive differential delay conditioning paradigm in patients with BPD and healthy controls (HC). We additionally assessed levels of peri-experimental dissociation, i.e., disruptions of usually integrated functions of consciousness, memory, body awareness, and perception. In this previous study, BPD patients with low levels of dissociation showed regular conditioning responses in terms of a differential SCR to CS+ > CS− similar to HC. Of note, patients with high state dissociation failed to show regular conditioning, which suggests that dissociation may modulate emotional learning in BPD.

Recently, Kamphausen and colleagues [43] assessed SCR and brain activity during an instructed fear task in patients with BPD and HC during functional magnetic resonance imaging (fMRI). Before participants were brought to the MR scanner, they were informed that they would see different stimuli either indicating a safe situation or potential threat (electrodermal stimulation). The electrodermal stimulation was in fact only applied during this instruction but not during the experimental procedure. During fMRI, both healthy participants and BPD patients showed differential fear responses in terms of stronger SCR to stimuli indicating potential threat compared to stimuli indicating the save situation. However, only HC but not BPD patients showed amygdala habituation and an increase in activity in the mPFC during instructed fear. Findings of this fMRI study suggest that altered activity in a fronto-limbic brain circuitry might underlie disturbed emotional learning during an instructed fear task in patients with BPD.

The aim of the present study was to investigate the neural correlates of emotional learning in BPD patients applying an aversive differential (danger versus safety signal) delay conditioning paradigm during fMRI. Between-session extinction and re-acquisition were assessed after 72 h. Based on previous neuroimaging studies, we expected stronger activation in the amygdala and insula during acquisition and reacquisition and less activation in the mPFC during within-session and between-session extinction. We were further interested in habituation effects, i.e., decreases in activity over the course of each conditioning phase. Based on previous research [7, 8, 43], we expected a habituation of the amygdala during the acquisition phase.

Methods and materials

Participants

Thirty-four female individuals fulfilling DSM-IV criteria for BPD as assessed by the international personality disorder examination (IPDE) [44] and 32 female HC aged between 18 and 45 years participated in our study. Exclusion criteria for the patient group were a lifetime diagnoses of alcohol or substance dependence, psychotic disorder or bipolar I disorder, and current major depression as assessed by the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I) [45]. In addition, patients had to be free of psychotropic medication for at least 4 weeks prior to investigation. Exclusion criteria for the HC group were the presence of any current or lifetime axis I or axis II disorder, and current or past psychotherapy. Diagnostic interviews were administered by trained and experienced psychologists. Inter-rater reliability for BPD was $\kappa = 0.69$ for SCID-I (primary diagnosis) and $\kappa = 0.77$ for IPDE. Patients were recruited at the Central Institute of Mental Health, Department of Psychosomatic Medicine and Psychotherapy in Mannheim/Germany or via announcements on BPD-associated Web sites. The HCs were randomly selected from the resident register of the city of Mannheim.

All participants underwent diagnostic assessment using the SCID-I [45] and the IPDE [44]. To obtain a psychopathological characterization of the patient sample, trait dissociation by the German version of the Dissociative Experience Scale (Fragebogen zu Dissoziativen Symptomen, FDS; [46]) and global BPD symptom severity by the Borderline Symptom List (BSL-95; [47]) were assessed.

Demographic and clinical characteristics (trait dissociation, BPD symptom severity, comorbid diagnoses) of the full sample can be found in Supplemental Table 1. There were no significant differences regarding age [BPD (Mean ± SD): 27.6 ± 7.5; HC: 27.8 ± 7.5], sex (all female), and ethnicity (all Caucasian). All participants were right-handed. There was a statistical trend for significant differences in level of education. As expected, both groups differed significantly in clinical measures of trait dissociation and BPD symptom severity.

From this initial sample, one HC and five BPD patients dropped out of the study before completion of the experiment. From the remaining sample, a subset of neuroimaging data had to be discarded due to movement artifacts or technical problems during MR scanning. The final neuroimaging data set comprised data from 26 HC and 27 BPD patients for the first (conditioning) session, 22 HC and 21 BPD patients for extinction recall, and 17 HC and 24 BPD patients for reacquisition. Since some participants did not show an adequate skin conductance response (SCR) > 1 μS during the scanning procedure, SCR data of 21 HC and 21 BPD from the first (conditioning) session and SCR of 15 HC and 14 BPD from the second session (extinction recall, reacquisition) could be included into the final psychophysiology analysis.

Demographic and clinical characteristics (trait dissociation, BPD symptom severity) of all subsamples can be found in Supplemental Table 2 (subsamples neuroimaging dataset) and Supplemental Table 3 (subsamples SCR dataset). There were no significant differences regarding age and educational level in any subsample. As expected, within all subsamples, BPD patients reported significantly more trait dissociation and BPD symptom severity.

Experimental design

The experimental design was an aversive differential delay conditioning procedure which is depicted by Fig. 1. Dependent variables were the blood-oxygen-level-dependent (BOLD) signal, SCR, and valence/arousal ratings. The conditioned stimuli (CS) were two neutral graphic patterns (blue square and yellow triangle). The US consisted of unpleasant, but tolerable electrical stimulation to the right thumb. Stimulus duration was 5.8 s (s), and the unconditioned stimulus (US) lasted for 2.8 s and was administered 3 s after CS onset. The inter-trial interval (ITI) lasted between 8 and 12 s, randomly generated. CS+ (CS paired with the US) and CS− (non-reinforced CS) were counterbalanced across participants and presented in pseudorandom order with the constraint of a maximum of two consecutive presentations of the CS+ (danger signal) or CS− (safety signal).

The US (electrical stimulation) was delivered through a copper electrode of an electrical stimulus generator (Digitimer, DS7A, Wewyn Garden City, UK). Before the experiment, for each participant, the level of pain stimulation was individually determined. To this end, participants were given a series of stimuli, starting with a mild stimulus, which was gradually increased to (1) detection threshold, (2) pain threshold, and (3) pain tolerance. This procedure was repeated three times. Finally, we delivered stimuli that were 80 % above pain threshold. Participants were asked to rate the intensity of pain on a Likert scale ranging from 0 = not painful to 10 = extremely painful. For the experiment, we used stimulation intensities that were rated with at least 7/10. Therefore, the experimental level of pain was objectively different, but subjectively both groups received the same level of painful stimulation. This procedure was chosen due to the fact of higher pain threshold in BPD [48, 49].

The first session (conditioning session) started with a phase in which participants were familiarized with the CS+ and CS− (*'familiarization'*: 10 CS+, 10 CS−). During the following *acquisition* phase, 18 CS+ paired with the US

Table 1 Brain activity during the different conditioning phases in healthy controls (HC) and in patients with borderline personality disorder (BPD)

Conditioning phase	Group	T contrast	MNI coordinates (*X*, *Y*, *Z*)	*K*	Hemisphere	Label (Brodmann area)	Lobe	*T* value	*Z* value	*p* value
Early acquisition	HC	CS+	0, −66, 12	110	R	Posterior cingulate	Limbic	6.12	4.71	0.037
			−27, −48, −15	49	L	Fusiform gyrus (BA37)	Occipital	6.25	4.77	0.021
			−18, −9, −12*	53	L	Parahippocampal gyrus/uncus/amygdala	Limbic	5.33	4.03	<0.05, ROI
			−24, −3, −24*		L		Limbic	4.83	3.78	
			18, −15, −15*	33	R	Parahippocampal gyrus (BA28)	Limbic	4.83	3.78	<0.05, ROI
		CS+ > CS−	No significant clusters at p(FDR) < 0.05							
	BPD	CS+	No significant clusters at p(FDR) < 0.05							
		CS+ > CS−	33, 39, 15	178	R	Middle frontal gyrus	Frontal	7.26	5.28	0.003
			27, 51, 3		R	superior frontal gyrus	frontal	4.17	3.60	0.003
			−60, 3, 3	49	L	Superior temporal gyrus	Temporal	5.29	4.30	0.022
			−66, −27, 12	61	L	Superior temporal gyrus (BA42)	Temporal	4.90	4.07	0.035
			66, −33, 18	89	R	Superior temporal gyrus (BA42)	Temporal	4.84	4.03	0.036
			−33, 15, 3*	16	L	Insula	Sub-lobar	4.32	3.70	<0.05, ROI
			−39,−3,−9*		L			3.63	3.22	<0.05, ROI
Late acquisition	HC	CS+	48, 33,−6*	34	R	Inferior frontal gyrus (BA47)	Frontal	4.80	4.00	<0.05, ROI
		CS+ > CS−	−30, −66, −51	1716	L	Inferior semilunar lobule	Cerebellum	6.32	4.84	0.004
			−33, −78, −30			declive	posterior	6.14	4.75	0.004
			−39, −63, −42			cerebellar tonsil		5.91	4.63	0.004
			45, −48, 51	606	R	Inferior parietal lobule (BA40)	Parietal	6.16	4.76	0.004
			63, −42, 24					5.59	4.46	0.004
			48, −45, 30					5.35	4.33	0.004
			54, 21, 33	178	R	Middle frontal gyrus (BA46)	Frontal	5.62	4.48	0.004
			54, 30, 18			inferior frontal gyrus (BA46)	frontal	4.27	3.67	0.008
			42, 6, 57			inferior frontal gyrus		4.26	3.66	0.008
			−60, −45, 24	216	L	Inferior parietal lobule (BA40)	Parietal	5.29	4.29	0.004
			−42, −51, 33					4.51	3.82	0.006
			−33, −51, 36					4.38	3.74	0.007
			39, 42, 30	35	R	Middle frontal gyrus	Frontal	4.95	4.10	0.005
			48, 27, −9*	85	R	Inferior frontal gyrus (BA47)	Frontal	4.93	4.08	<0.05, ROI
			−54, 21, 0*	41	L	Inferior frontal gyrus (BA47)	Frontal	4.22	3.63	<0.05, ROI
			−51, 15, −6*				frontal	4.16	3.59	<0.05, ROI
	BPD	CS+	48, 15, 45	49	R	Middle frontal gyrus (BA8)	Frontal	6.76	5.09	0.011
			39, 57, 3	98	R	Middle frontal gyrus (BA8)	Frontal	5.88	4.65	0.027
			60, −24, −24	32	R	Inferior temporal gyrus	Temporal	5.36	4.36	0.028
			−27, −66, −30	158	L	Pyramis	Cerebellum	5.15	4.24	0.030
			−57, −36, −6	38	L	Middle temporal gyrus	Temporal	4.83	4.04	0.040
			6, 39, 42	87	R	Medial frontal gyrus (BA8)	Frontal	5.06	4.18	0.040

Table 1 continued

Conditioning phase	Group	T contrast	MNI coordinates (*X*, *Y*, *Z*)	*K*	Hemisphere	Label (Brodmann area)	Lobe	*T* value	*Z* value	*p* value
		CS+ > CS−	48, 15, 45	208	R	Middle frontal gyrus (BA8)	Frontal	7.08	5.55	0.001
			48, 39, 30		R	Middle frontal gyrus (BA46)	frontal	4.44	3.80	0.017
			−18, −102, 3	831	L	Middle occipital gyrus (BA18)	Occipital	6.29	4.86	0.004
			39, 57, 3	120	R	Middle frontal gyrus (BA9)	Frontal	5.99	4.71	0.004
			30, 12, 3	46	R	Inferior frontal gyrus	Frontal	4.93	4.11	0.012
			33, 21, −9		R			3.89	3.42	0.027
			60, −42, 42	58	R	Inferior parietal lobule (BA40)	Parietal	4.51	3.84	0.016
			6, 36, 42	32	R	Medial frontal gyrus (BA8)	Frontal	4.44	3.79	0.017
			−45, 54, −3	45	L	Middle frontal gyrus (BA10)	Frontal	4.39	3.77	0.018
Extinction	HC	CS+	48, 48, −3*	R	18	Middle frontal gyrus	Frontal	5.57	4.45	<0.05, ROI
			−39, 15, 45*	L	31	Middle frontal gyrus	Frontal	5.01	4.13	<0.05, ROI
			9, 36, 24*	R	57	Orbito frontal gyrus	Frontal	4.86	4.04	<0.05, ROI
			33, 63, 18*					4.63	4.63	<0.05, ROI
			21, 63, 21*					3.87	3.87	<0.05, ROI
			−45, 39, −9*	L	20	Inferior frontal gyrus (BA 47)	Frontal	4.54	3.84	<0.05, ROI
			−27, 57, 24*	L	15	Middle frontal gyrus	Frontal	4.23	3.64	<0.05, ROI
			45, 48, 27*	R	14	Middle frontal gyrus	Frontal	4.13	3.57	<0.05, ROI
			51, 45, 21*	R		middle frontal gyrus	frontal	4.12	3.56	<0.05, ROI
			42, 36, 24*	R	9	Middle frontal gyrus	Frontal	4.07	3.53	<0.05, ROI
		CS+ > CS−	24, 66, 21*	R	28	Superior frontal gyrus (BA10)	Frontal	3.96	3.45	<0.05, ROI
			15, 63, 24*	R		Superior frontal gyrus	frontal	3.90	3.41	<0.05, ROI
			−45, 21, 45*	L	7	Middle frontal gyrus	Frontal	3.83	3.36	<0.05, ROI
	BPD	CS+	45, 42, −12*	R	99	Orbitofrontal gyrus (BA11)	Frontal	4.76	4.00	<0.05, ROI
			30, −6, −12*	R	5	Amygdala	Limbic	3.90	3.43	<0.05, ROI
		CS+ > CS−	No significant clusters p(FDR) < 0.05							
Extinction recall	HC	CS+	12, −63, 0	2235	R	Lingual gyrus	Occipital	10.14	6.12	<0.001
			6, −78, 15		R	lingual gyrus	Occipital	9.14	5.82	<0.001
			−12, −84, 3		L	cuneus (BA 17)	occipital	7.16	5.09	<0.001
			−36, 42, −9	17	L	Middle frontal gyrus	Frontal	5.00	4.04	0.002
			−39, −63, −39	31	L	Tuber	Cerebellum	4.79	3.92	0.003
			51, −48, 3	33	R	Middle temporal gyrus	Temporal	3.80	3.30	0.016
			63, −51, −9		R	middle temporal gyrus (BA21)	Temporal	3.76	3.27	0.017
			−27, −3, −27	10	L	Parahippocampal gyrus	Limbic	4.38	3.67	0.006
			−24, 51, 39	4	L	Superior frontal gyrus (BA9)	Limbic	3.86	3.34	0.014
			−36, −24, −15	6	L	Parahippocampal gyrus	Limbic	3.80	3.30	0.016
		CS+ > CS−	No significant clusters $p < 0.05$							

Table 1 continued

Conditioning phase	Group	T contrast	MNI coordinates (*X*, *Y*, *Z*)	*K*	Hemisphere	Label (Brodmann area)	Lobe	*T* value	*Z* value	*p* value
	BPD	CS+	12, −72, 9	449	R	Cuneus (BA17)	Occipital	6.78	4.78	0.032
			45, 9, −9*	49	R	Insula (BA38)	Sub-lobar	4.34	3.58	0.05, ROI
			−27, −3, −27	3	L	Parahippocampal gyrus	Limbic	4.38	3.67	0.009
		CS+ > CS−	24, 39, −18*	7	R	Orbitofrontal gyrus (BA11)	Frontal	4.90	3.89	<0.05, ROI
			6, 18, 51*	2	R	medial frontal gyrus (BA8)	Frontal	4.00	3.37	<0.05, ROI
Re-acquisition	HC	CS+	51, −45, 18	13705	R	Superior temporal gyrus,	Temporal,	11.53	6.01	<0.001
			−51, −24, 24		L	insular cortex (BA13)	Sub-lobar	11.09	5.91	<0.001
			−66, −27, 30		L			9.78	5.60	<0.001
			48, 12,0*					8.88		<0.05, ROI
			30, 57, 18	219	R	Superior frontal gyrus/middle frontal gyrus (BA9,	Frontal	7.78	5.02	<0.001
			27, 51, 24		R	BA10)	Frontal	6.88	4.70	<0.001
			42, 54, 12		R		Frontal	6.22	4.43	<0.001
					R					
			−39, 45, 24	74	L	Middle frontal gyrus (BA10, BA9)	Frontal	4.94	3.84	0.001
			−27, 51, 30		L			4.90	3.82	0.001
			−48, 45, 15		L			4.74	3.73	0.001
			−30, 3, −18*	45	L	Parahippocampal gyrus (BA34)	Limbic	5.10	3.92	<0.05, ROI
			−24, 0, −24*		L	amygdala/uncus	Limbic	5.14	3.94	<0.05, ROI
		CS+ > CS−	−3, 51, −6*	151	L	Medial frontal gyrus (BA10)	Frontal	6.80	4.28	<0.05, ROI
			0, 39, −18*		R			5.67	3.88	<0.05, ROI
			12, 45, −12*		R			4.98	3.60	<0.05, ROI
			−21, 42, 45*	L	19	Superior frontal gyrus	Frontal Frontal	5.18	3.68	<0.05, ROI
			−15, 42, 39*			superior frontal gyrus		5.03	3.62	<0.05, ROI
			−27, 24, 51*	10	L	Middle frontal gyrus (BA8)	Frontal	4.59	3.42	<0.05, ROI
	BPD	CS+	9, 3, 39*	19*	R*	Anterior cingulate (BA24)	Limbic	4.64	3.84	<0.05, ROI
			39, 54, −15*	16*	R*	Orbitofrontal gyrus (BA11)	Frontal	4.38	3.67	<0.05, ROI
		CS+ > CS−	36, 51, 12*	10	R	Middle frontal gyrus (BA10)	Frontal	4.28	3.61	<0.05, ROI

Clusters were determined applying an extended threshold of $p < 0.05$ FDR corrected on the whole-brain voxel-wise level. Clusters indicated by an *asterisk* were detected by region of interest (ROI) analyses with predefined anatomical masks. Sizes of subsamples were 26 for session 1 (familiarization, acquisition, extinction), 22 for extinction recall, and 17 for re-acquisition. *Note L* left, *R* right, *k* cluster size of contiguous voxels, *MNI* Montreal Neurological Institute, *R/L* right/left, *BA* Brodmann area, *CS+* conditioned stimulus that was sometimes paired with the unconditioned stimulus (US) but here only in absence of the US, *CS−* conditioned stimulus never paired with US

and 36 CS− never paired with the US were presented. Furthermore, 18 CS+ were presented but not paired with the US (catch trials). Catch trials were used to differentiate between BOLD responses to the CS+ and to the pain stimulus (US). During the *extinction* phase, 18 CS+ and 18 CS− were presented. The second session (reacquisition session) took place 72 h later and consisted of the *extinction recall* phase (18 CS+ and 18 CS−) and a *reacquisition* phase (18 CS−, 9 CS+, 9 CS+/US).

Apparatus and physiological recordings

Stimulus delivery was controlled by a PC running Presentation software (Neurobehavioral Systems, San Francisco). Physiological data acquisition was controlled by an Ibook running Vitagraph version 4.61 (Becker Meditec, Karlsruhe, Germany). Physiological channels were recorded at a rate of 256 Hz in continuous mode using the Vitaport II system (Becker). SCR was obtained from 10-mm (sensor diameter) Ag/AgCl electrodes (Marquette Hellige GmbH, Freiburg, Germany) filled with an isotonic EDR jelly TDE-246 (Steffens, Berlin, Germany) and placed on the thenar and hypothenar of the non-dominant hand (constant voltage method with 0.5 V).

Procedure

The study was approved by the ethical review committee of the University of Heidelberg, Germany, in accordance with the Declaration of Helsinki. All subjects gave written informed consent after receiving a description of the study and scanning procedure. All participants were paid for study participation (10 €/h). The experiment was conducted between 2007 and 2009 at the Central Institute of Mental Health in Mannheim, Germany.

All participants underwent diagnostic assessments including the SCID-I and IPDE by trained diagnosticians and completed clinical questionnaires (FDS, BSL) as described above. The conditioning procedure took place on two different days within a fixed time interval. On a first day, the conditioning session (familiarization phase, acquisition phase, extinction phase) was conducted during fMRI. Immediately before the experiment, the level of pain stimulation was individually determined for each participant as described in detail above. The reacquisition session (extinction recall phase and reacquisition phase) took place 72 h later in the same MR scanner.

Throughout the scanning procedure, participants viewed stimuli presented on a screen by a projector via a mirror mounted on the head coil. Participants were not informed about CS–US contingencies and were told to passively view the stimuli. After each phase (familiarization acquisition, extinction, extinction recall, and reacquisition), participants rated the valence and arousal of the CSs to test awareness of the CS–US contingency. In addition, participants rated their acute dissociation on scales ranging from 1 = very calm to 9 = very arousing, 1 = very pleasant to 9 = very unpleasant, and 1 = no dissociative symptoms to 9 = very intense dissociative symptoms. Valence/arousal ratings of the CS were obtained by presenting the CS together with a visual analogue scale, and patients used a keypad to move the cursor and to choose a number.

fMRI assessment

fMRI was conducted by a Siemens MAGNETOM Vision 1.5 Tesla whole body scanner (Siemens Medical Solution, Erlangen, Germany). Head movement artifacts and scanning noise were restricted using head cushions and headphones within the scanner coil. To acquire BOLD signals (T2-weighted echo planar imaging, EPI), a standard protocol (see also [27]) with the following parameters was used: repetition time (TR): 3.77 s, echo time (TE): 45 ms; flip angle: 90°, matrix: 64 × 64, slice thickness: 3 mm, slice gap: 1 mm, FOV: 220 × 220 mm, and number of slices: 35. There were 100 images for familiarization, 320 for acquisition, and 160 for all other phases.

Data analysis

Psychometrics

Means and standard deviations of the BSL and the FDS were analyzed using t tests. For group comparisons, significance levels were set to $p < 0.05$. All analyses were conducted with SPSS (Version 22.0 for Windows; SPSS Inc., Chicago, IL).

Ratings

To compare valence and arousal ratings, repeated-measures analyses of covariance (rm-ANCOVA) were performed for the mean of each phase with group (HC, BPD) as between-subject factor, stimulus type (CS+, CS) as within-subjects factor, and present-state dissociative experience as covariate. A Greenhouse-Geisser correction was employed in case the sphericity assumption was not met. Statistical analysis was performed using SPSS 22.0 (SPSS Inc, Chicago, IL). An alpha level of $p < 0.05$ determined statistical significance.

Skin conductance response (SCR)

The electrodermal activity was determined according to published guidelines [50] with the program EDR-Para (Schäfer, Wuppertal). Phasic SCRs were defined as the response

magnitude (maximum deflection) within a 1- to 4-s time frame (first interval response, FIR). SCRs lower than 0.05 μS were scored as zero. The FIR was counted as missing when the SCR was clearly initiated prior to CS onset. If such an anticipatory response was superimposed on a stimulus-related response, the scoring method (B) as described by Boucsein ([50], p. 136) was used. SCRs were log-transformed to reduce skewness [50]. To compare SCR to CS+ and CS− type × group, rm-ANCOVAs were performed for the mean of each phase with present-state dissociative experience as covariate. A Greenhouse-Geisser correction was employed if the sphericity assumption was not met. Statistical analysis was performed using SPSS 22.0 (SPSS Inc, Chicago, IL). An alpha level of <0.05 determined statistical significance.

fMRI data

Functional data were analyzed with SPM5 (Welcome Department of Imaging Neuroscience, University College London, UK, 2005). The first four images at the beginning of each trial were discarded to enable the signal to achieve steady-state equilibrium between radiofrequency pulse and relaxation.

Within preprocessing, the EPI time series were realigned and unwarped to the mean to correct for intra-subject's head movements. Mean images were normalized to an MNI (Montreal Neurological Institute, www.bic.mni.mcgill.ca) echoplanar imaging template with affine registration followed by nonlinear transformation with 25 mm cutoff, medium regularization and 16 iterations, resampled with trilinear interpolation, and written in 3 × 3 × 3 mm^3 isotropic voxels. The normalization parameters determined for the mean functional volume were then applied to the corresponding functional image volumes for each participant. Finally, images were smoothed with a Gaussian kernel of 8 mm full-width at half-maximum. The data were high-pass filtered (1/128 Hz cutoff) to remove low-frequency signal drifts.

The consecutive statistical analyses of the fMRI data relied upon the general linear model (GLM) to model effects of interest as implemented in SPM5. For each subject, the following events of interest were defined as regressors of interest: US, CS−, $CS+_{paired}$, and $CS+_{unpaired}$ for each phase of the conditioning paradigm: familiarization, acquisition, extinction, extinction recall, and reacquisition phase.

In order to assess the temporal course of brain activation, we split up the acquisition phase (of the first conditioning session) into an early and late acquisition phase. In addition, we investigated habituation effects during acquisition by modeling linear time effects for $CS+_{unpaired,}$ $CS+_{unpaired}$ > CS−, and $CS+_{paired}$ (CS paired with US) using the time modulation function as implemented in SPM ($p < 0.05$).

Task-related activity (BOLD response) for each event was modeled by convolving a vector of the event onset times with a canonical hemodynamic response function (HRF). The regressors of interest were included together with time as covariate in a GLM to yield parameter estimates for brain activation related to the different event types (US, CS−, $CS+_{paired}$, $CS+_{unpaired}$). T contrast images for $CS+_{paired}$ > CS− and CS− > $CS+_{paired}$ were computed for each conditioning phase (familiarization, early acquisition, late acquisition, extinction, extinction recall, reacquisition) separately.

At the second level (group level), the resulting first level contrast images were entered into random effect models: *Within-group* (one sample) *t* tests were used to compare brain activity during CS+ to brain activity in response to CS−. *Between-group* (two sample) *t* tests were used to analyze group differences in response to CS+, CS−, as well as $CS+_{unpaired}$ > CS− during the different conditioning phases. Clusters were determined using an extent threshold of $p < 0.05$ on the voxel-wise whole-brain level. The false discovery rate (FDR) correction was applied to correct for multiple comparisons.

Region of interest analyses Based on our a priori hypotheses (stronger activation in the amygdala and insula during acquisition and reacquisition and less activation in the mPFC during within-session and between-session extinction in BPD patients than HC), regions of interest analyses (ROIs) were performed for the following brain areas: amygdala, ACC, hippocampus, insula, OFC, dfPFC, and vmPFC. Anatomically based ROIs of these regions were created using the Masks for Regions of Interest Analysis (MARINA) software program (Bertram Walter, Bender Institute of Neuroimaging, University of Giessen, Germany). These masks were then used for small volume corrections.

In a separate model, state dissociation ratings were included as a covariate in order to test the effect of this

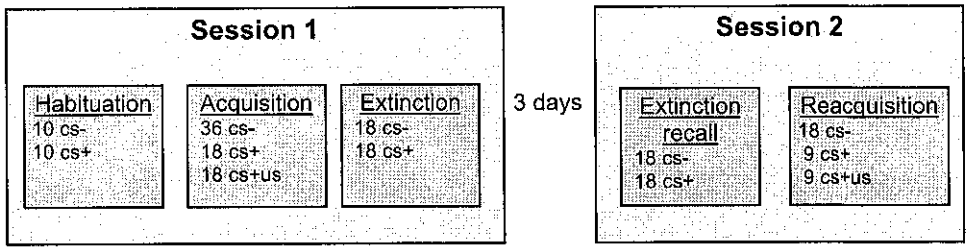


Fig. 1 Experimental design: Ratings of valence and arousal were performed after each of the five experimental phases. *CS+* conditioned stimulus paired with the US; *CS−* non-reinforced conditioned stimulus, *US* unconditioned stimulus

variable on the results. Furthermore, a subgroup analysis with BPD patients with versus without current PTSD compared to HC was computed.

Results

Ratings

Means and standard deviations of arousal (a) and valence (b) ratings during acquisition, extinction, extinction recall, and reacquisition in HC and BPD patients are depicted in Fig. 2.

For arousal ratings after acquisition, a significant effect of type [$F(1,34) = 23.929$, $p < 0.001$] but no significant interaction with group was found [$F(1,34) = 0.001$, $p = 0.973$]: All participants showed a higher rating of the CS+ compared to the CS−. For the arousal ratings after extinction and extinction recall, no significant effect of type [$F(1,30) = 0.250$, $p = 0.621$; $F(1,24) = 0.566$, $p = 0.459$] and no significant interaction with group were found [$F(1,30) = 0.009$, $p = 0.925$, $F(1,24) = 0.452$, $p = 0.508$]. After extinction, BPD patients rated both stimuli as more aversive than the HC indicated by a significant effect of group [$F(1,30) = 5.565$, $p = 0.025$].

For the arousal ratings after reacquisition, a significant effect of type [$F(1,24) = 18.315$, $p < 0.001$] and no significant interaction with group were found [$F(1,24) = 0.337$, $p = 0.567$]: All participants showed a more aversive rating of the CS+ compared to the CS− (see Fig. 2a).

For the valence ratings after acquisition, a significant effect of type [$F(1,34) = 15.199$, $p < 0.001$], with more aversive rating of the CS+ than the CS−, and a significant interaction effect type × group [$F(1,34) = 4.223$, $p = 0.048$] were found. For the valence ratings after extinction and extinction recall, no significant effect of type [$F(1,30) = 0.334$, $p = 0.568$; $F(1,24) = 1.189$, $p = 0.286$] and no interaction with group were found [$F(1,30) = 0.021$, $p = 0.886$, $F(1,24) = 1.156$, $p = 0.293$]. After extinction, again, a significant effect of group was found for the valence ratings [$F(1,30) = 14.944$, $p = 0.001$]: BPD patients rated both stimuli as more aversive than the HC. For the valence ratings after reacquisition, a significant effect of type [$F(1,24) = 14.138$, $p = 0.001$] but no interaction with group was found [$F(1,24) = 0.199$, $p = 0.660$]: All participants showed a more aversive rating of the CS+ compared to the CS− (see Fig. 2b).

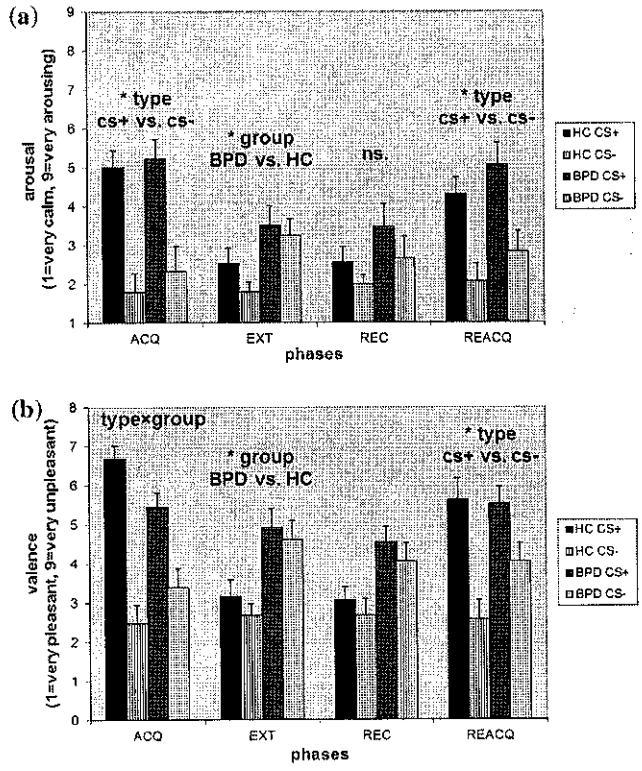


Fig. 2 Results of the arousal (**a**) and valence (**b**) ratings. *ACQ* acquisition, *EXT* extinction, *REC* extinction recall, *REACQ* reacquisition, *HC* healthy controls, *BPD* BPD patients, *CS+* conditioned stimulus paired with the US, *CS−* non-reinforced conditioned stimulus, significant at $^{*}p < 0.05$, $^{\#}p < 0.10$

Skin conductance response (SCR)

Means and standard deviations of SCR during acquisition, extinction, extinction recall, and reacquisition in HC and BPD patients are shown in Fig. 3. During acquisition, no significant effect of type [$F(1,34) = 2.291$, $p = 0.139$, $\eta^2 = 0.063$] and no significant interaction with group were found [$F(1,34) = 0.039$, $p = 0.844$]. During extinction, no significant interaction with group was found [$F(1,33) = 0.962$, $p = 0.334$], but a significant effect of type [$F(1,33) = 9.895$, $p = 0.003$] with participants shows stronger responses to the CS+ than to the CS−. During extinction recall, no significant effect of type [$F(1,20) = 0.209$, $p = 0.652$] and a trend for the interaction effect type x group were found [$F(1,20) = 3.194$, $p = 0.089$], with BPD patients showing descriptively a stronger response to CS+ compared to the CS−, whereas HCs showing descriptively a stronger response to CS− compared to the CS+. During reacquisition, no significant effect of type [$F(1,20) = 1.967$, $p = 0.176$, $\eta^2 = 0.090$] and no interaction with group were found [$F(1,20) = 1.768$, $p = 0.199$, $\eta^2 = 0.081$].

fMRI

Between-group t tests

No significant group differences on the whole-brain level could be detected at a threshold level of $p < 0.05$ after correction for multiple comparisons (FDR correction).

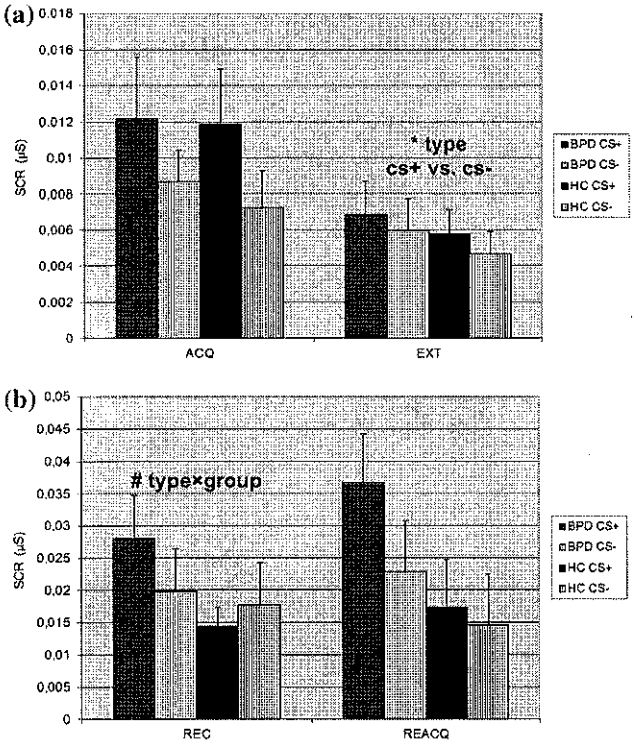


Fig. 3 Skin conductance response. *Note ACQ* acquisition, *EXT* extinction, *REC* extinction recall, *REACQ* reacquisition, *HC* healthy controls, *BPD* BPD patients, *CS+* conditioned stimulus paired with the US, *CS−* non-reinforced conditioned stimulus

Within-group t tests

Results for brain activity in response to $CS+_{unpaired}$ and $CS+_{unpaired} > CS-$ can be found in Table 1. In the following, results are reported for each group and for each conditioning phase separately [p(FDR) < 0.05].

Healthy control group

During *early acquisition*, no significant clusters were observed for the $CS+_{unpaired} > CS-$ contrast. During presentation of the $CS+_{unpaired}$, HC recruited the fusiform gyrus (BA37) and right posterior cingulate. Our ROI analyses further revealed significant $CS+_{unpaired}$-related activation in the bilateral parahippocampal gyrus and left amygdala.

During *late acquisition*, HC showed significant stronger activation in response to the $CS+_{unpaired}$ than to CS− in the bilateral inferior frontal gyrus (BA46 and BA47), bilateral inferior parietal lobule, right middle frontal gyrus (BA46), and left cerebellum.

During *extinction*, HC exhibited significant differential activation for $CS+_{unpaired} > CS-$ in the right superior frontal gyrus (BA10) and left middle frontal gyrus (clusters revealed by ROI analyses). In addition, our ROI analyses revealed significant $CS+_{unpaired}$-related activation in the right OFC, left inferior frontal gyrus (BA47), and bilateral middle frontal gyrus.

During *extinction recall*, no significant clusters were found for the $CS+_{unpaired} > CS-$ contrast. In response to the $CS+_{unpaired}$, HC activated the right lingual gyrus and left cuneus, the left middle frontal gyrus and superior frontal gyrus (BA9), the left parahippocampal gyrus, right middle temporal gyrus, and left tuber (cerebellum).

During *reacquisition*, our ROI analyses revealed significant clusters in the bilateral medial frontal gyrus (BA10), left superior frontal gyrus, and left middle frontal gyrus (BA8) for the $CS+_{unpaired} > CS-$ contrast. In response to the $CS+_{unpaired}$, HC showed recruitment of the bilateral superior temporal gyrus and insular cortex (BA13), bilateral middle frontal, and superior frontal. For the $CS+_{unpaired}$ contrast, ROI analyses additionally revealed significant clusters in the left amygdala and left parahippocampal gyrus.

BPD group

During *early acquisition*, BPD patients showed significant differential activation for $CS+_{unpaired} > CS-$ in the right superior and middle frontal gyrus as well as in the bilateral superior temporal gyrus. ROI analyses further revealed significant stronger activation in the left insula (see Fig. 4a) for the $CS+_{unpaired} > CS-$ contrast.

During *late acquisition*, BPD patients exhibited significant stronger activation in response to $CS+_{unpaired} > CS-$ in the bilateral middle and medial frontal gyrus (BA8, BA9, BA10, BA46), right inferior frontal gyrus, left middle occipital gyrus (BA18), and right inferior parietal lobule (BA40). In response to the $CS+_{unpaired}$, in addition, patients recruited the right inferior temporal gyrus and left middle temporal gyrus.

During the *extinction* phase, no significant clusters were observed for the $CS+_{unpaired} > CS-$ contrast. In response to $CS+_{unpaired}$, BPD patients activated the right orbitofrontal gyrus (BA 11) and right amygdala (see Fig. 4b).

During *extinction recall*, BPD patients exhibited $CS+_{unpaired}$-related activation in the right cuneus and left parahippocampal gyrus as well as in the right insula. For the $CS+_{unpaired} > CS-$ contrast, ROI analyses revealed stronger activation in the orbitofrontal gyrus (BA11) as well as medial frontal gyrus (BA8) in the patient group.

During *reacquisition*, we found significant clusters in the middle frontal gyrus (BA10) for the $CS+_{unpaired} > CS-$ contrast as well as clusters in the orbitofrontal gyrus (BA11) and ACC (BA24) for the $CS+_{unpaired}$ contrast.

Linear habituation

Complete results for linear time effects during the acquisition phase in each group can be found in Supplemental Table 4.

Healthy control group

In HC, we found a habituation effect, i.e., significant decrease in activity in the mid-cingulate gyrus in response to $CS+_{unpaired}$ as well as $CS+_{unpaired} > CS-$. Our ROI analysis further revealed a significant linear decrease in activity in the right amygdala and bilateral parahippocampal gyrus for the repeated exposure to the CS in temporal contingency with the US ($CS+_{paired}$) (see Supplemental Table 4).

BPD group

BPD patients showed a habituation of left insula activation and left medial frontal gyrus activation in response to $CS+_{unpaired}$ as well as $CS+_{unpaired} > CS-$ during acquisition. For the $CS+_{paired}$ (i.e., repeated exposure to the CS in temporal contingency with the US), we found a significant linear decrease in activation in the left parahippocampal gyrus and inferior frontal gyrus (see Supplemental Table 4). The ROI analysis with the amygdala mask revealed no significant clusters.

Subgroup analysis

Analyses with state dissociation as a covariate and for the subgroup of BPD patients with comorbid PTSD led to similar results (data not shown).

Discussion

In this fMRI study, we investigated the neural correlates of emotional associative learning in un-medicated female patients with BPD compared to female HC participants. During an aversive differential delay conditioning paradigm, brain activity (BOLD responses), subjective ratings, and SCR were assessed. In addition to within-session extinction, we investigated extinction recall (between-session extinction) and reinstatement, i.e., reacquisition of the initial CS–US contingency, after 72 h. We were further interested in the temporal graduation of brain activation. Therefore, we slit the initial acquisition phase into an '*early acquisition*' and a '*late acquisition*' phase and estimated habituation effects in terms of a linear decrease in brain activity during each conditioning phase.

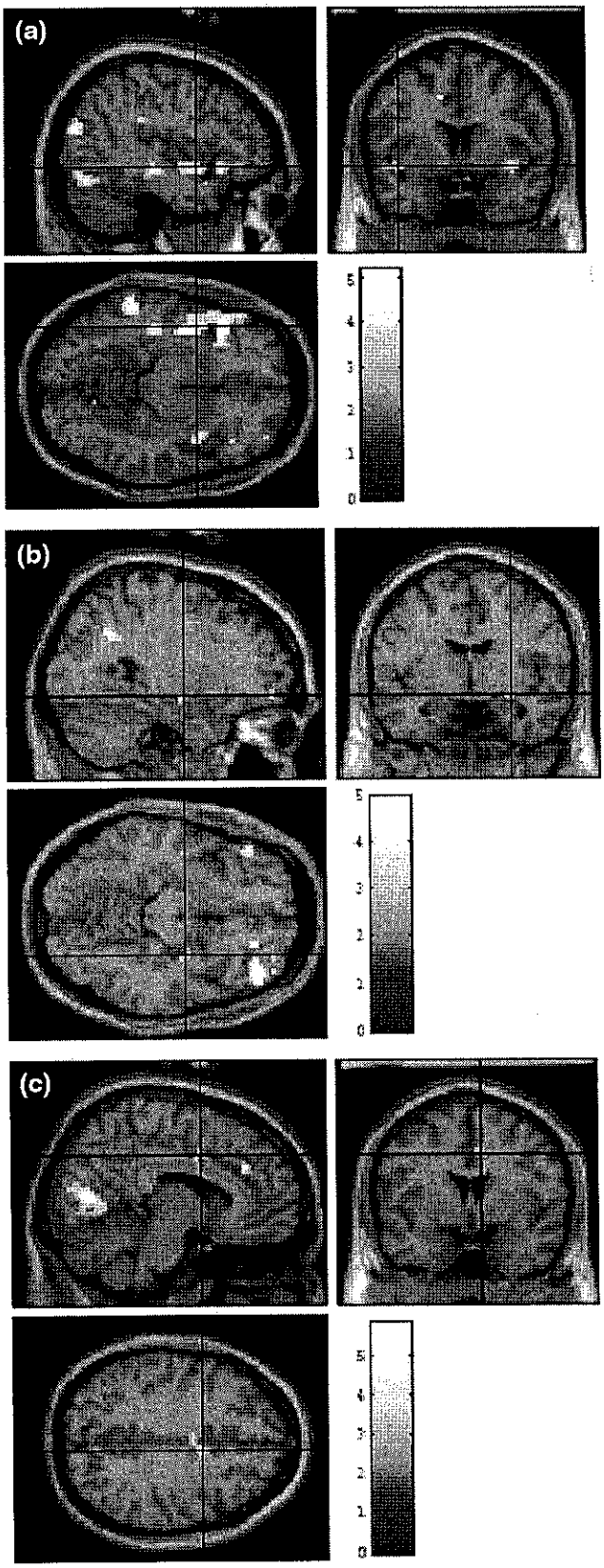
Concerning subjective ratings, both groups perceived the CS+ as more arousing than the CS− after acquisition as well as reacquisition, suggesting that they successfully learned the difference between safety and danger signal. During extinction, a significant group effect was found with higher arousal and valence ratings for CS+ but also for CS− in BPD patients than in healthy participants. This finding may point to disturbed emotional learning in terms of slower extinction processes in BPD patients compared to HC.

SCR was assessed as a more direct measure of arousal and a more proximal measure of the fear conditioning process. Differential learning effects by means of stronger responses to CS+ than to CS− were only observed for the extinction phase. However, since some participants did not show an adequate SCR (>1 μS), the sample size for our psychophysiology analysis was probably too small to detect statistically significant effects (as effect sizes were in a medium range). Interestingly, we found a trend for BPD patients responding stronger to CS+ than to CS− during extinction recall, again arguing for slower extinction learning in these patients.

On the neural level, no significant between-group differences were found after correcting for multiple comparisons. During acquisition as well as reacquisition, both groups showed significant $CS+_{unpaired}$-related activation in the amygdala, insula, and mPFC brain areas that have been identified as central parts of a 'classical fear network' (for a review see, e.g., [28, 31]).

Results of within-group comparisons and time effect analyses point to a different temporal course of brain activity in the two groups. During the early acquisition phase, BPD patients but not HC showed differential activation to $CS+_{unpaired} > CS-$ in the left insula. The insula has been implicated in external (e.g., detection of salient events including threat in the environment) as well as internal attention (e.g., interoceptive awareness, monitoring of arousal states) [51, 52]. It has been proposed that the insula plays an important role in conveying a cortical representation of fear to the amygdala [20], particularly in the context of uncertain aversive events [12, 13, 23]. Early recruitment of the insula may therefore suggest increased vigilance to potential threat already at an early stage of the acquisition phase in BPD patients. During the late acquisition phase, presentation of $CS+_{unpaired}$ (compared to CS−) was associated with increased activation in fronto-parietal brain regions including the dlPFC, OFC, vlPFC (inferior frontal gyrus), and inferior parietal lobule in both groups. Involvement of these brain areas in acquisition processes has also been reported in previous fear conditioning studies [16, 17, 28, 31]. It has been proposed that the dlPFC may predominantly be active when aversive (threatening) stimuli are predictable during fear conditioning [31]. The present neuroimaging findings may therefore be related to a successful differential learning of safety and danger signals at a later stage of the acquisition phase in both groups.

In our present study, we were not only interested in within-session extinction but also in *extinction recall* (between-session extinction) after 72 h. During within-session extinction, both groups recruited the OFC when exposed to the CS+. BPD patients—but not



◀**Fig. 4** Brain activity during the different conditioning phases in BPD patients: **a** Insula activation was found during the early acquisition phase, **b** right amygdala activation was found during extinction, **c** right mid-cingulate activation was observed during reacquisition

HC—additionally showed $CS+_{unpaired}$-related activation in the amygdala, whereas healthy participants additionally recruited the inferior frontal gyrus and middle frontal gyrus. Activation in the amygdala is thought to occur as an orienting response to (especially self-relevant) salient stimuli and is assumed to serve the relay of emotionally relevant information to different processing channels [9, 21, 28]. During extinction, BPD patients in our study rated both the CS+ and CS− as significantly more arousing and aversive than HC. In this context, CS+-related amygdala activation may be related to altered differential learning in these patients. While healthy participants showed significantly stronger prefrontal activation in response to CS+ > CS− during extinction, BPD patients showed increased activation for CS+ > CS− only during extinction recall (i.e., during the second extinction session after 72 h). Moreover, during extinction recall, BPD patients but not HC showed $CS+_{unpaired}$-related activity in the insular cortex. As mentioned above, we found a trend for stronger SCR to CS+ than to CS− in the BPD group. In this context, the present neuroimaging findings may point to delayed extinction learning in BPD patients.

When the original contingency of the CS+ and US was renewed after 72 h, HC showed $CS+_{unpaired}$-related activation in the insular cortex, amygdala, and parahippocampal gyrus—regions critically involved in fear acquisition [28]. After extinction, reacquisition gives a renewed meaning to the CS+. Amygdala and insula might play an important role in mediating these kinds of contingency changes and in the building of new associations [22]. During reacquisition, BPD patients but not HC activated the ACC (BA24), which has been associated with the anticipation of pain and the integration of nociceptive input with memory [6]. Both groups during reacquisition increased activity in the middle frontal gyrus (BA10) in response to the $CS+_{unpaired}$ compared to the CS−, but HC additionally recruited parts of the dlPFC (BA8). As mentioned above, the dlPFC has been proposed to play an important role in regulatory processes including emotion down-regulation [32] as well as processing aversive stimuli that are predictable [31].

In the present study, we were further interested in the temporal course of brain activation, particularly in linear habituation effects during acquisition. Interestingly, in the patient group, we found a linear decrease in activity in the

left insula in response to $CS+_{unpaired} > CS-$. Since patients showed increased activity in this area (suggesting increased vigilance) to the $CS+_{unpaired}$ during early acquisition, they may have learned the association between the CS and US relatively early and therefore reduced their attention to the $CS+_{unpaired}$ over the course of the acquisition phase. In HC, we found a linear habituation effect for the mid-cingulate cortex (BA 5) to the CS+ during acquisition. Of note, amygdala habituation to $CS+_{paired}$ during acquisition was only found in HC but not in the BPD group. This finding is in line with a recent fMRI study by Kamphausen and colleagues [43] who applied an instructed fear task. In this study, healthy participants but not BPD patients showed a habituation (i.e., decrease in activity) in the amygdala and an increase in activity in the mPFC over the course of the instructed fear task. Moreover, BPD patients previously showed a failure to habituate to the repeated presentation of aversive pictures which was associated with prolonged amygdala activity and smaller increases in amygdala–insula connectivity [53, 54].

All in all, our findings may be related to disturbed emotional learning processes in patients with BPD. Early activation in the insula in response to $CS+_{unpaired} > CS-$ suggests that patients may react more sensitively to fear conditioning in terms of increased vigilance already at the beginning of the acquisition phase. Amygdala activity during extinction and insula activity along with stronger SCR to CS+ than CS− during extinction recall—as well as the lack of this activation in HC—may point to delayed extinction processes in BPD patients. However, further fMRI studies are needed to confirm and extend these findings and may also investigate conditioning processes in a social context.

There are several limitations of this study. First of all, we lost data since several participants did not show an appropriate SCR due to dropouts or due to technical problems. The sample size for the analysis of the SCR was probably too small to find significant effects. The partial reinforcement during acquisition (CS+ not always followed by US) may have led to psychophysiological responses that were less clear-cut. Regarding fMRI assessment, we used a standard protocol with a TR of 3.77 s which is relatively long given the duration of the US of 2.8 s. Moreover, although we excluded substance dependence, psychotic disorders, and current depression as comorbid conditions, it might well have been that other comorbid conditions such as antisocial personality disorder or anxiety disorders have influenced learning processes and its neural correlates in the present study. We tested the influence of comorbid PTSD, but found no differences in any of our dependent variables between patients with and without PTSD. A clinical assessment of depressive symptoms (e.g., the BDI) would have been helpful to further characterize our sample.

Despite these limitations, we believe that our data are of importance as, to our knowledge, this is the first neuroimaging study investigating the neural correlates of acquisition, (within-session) extinction, (between-session) extinction recall, and re-acquisition within a classical delay conditioning paradigm in BPD. The alterations in the time course of neural response patterns during the different learning phases as demonstrated here—in conjunction with the influence of dissociation on conditioning as demonstrated earlier [42]—might be relevant for the understanding of disturbed learning processes in this disorder and should be considered in the context of learning-based psychotherapy. Findings may further point to the necessity of stress and emotion regulation training in these patients.

Acknowledgments We thank the Deutsche Studienstiftung for support (fellowship for Jana Keibel-Mauchnik) and all subjects for participating in our experiment. This study was supported by the German Research Foundation (SFB 636/C2).

Conflicts of interest There are no conflicts of interest.

References

1. Davis M (1992) The role of the amygdala in conditioned fear. In: Aggleton JP (ed) vol 9. Wiley, New York, pp 255–306
2. Molchan SE, Sunderland T, McIntosh AR, Herscovitch P, Schreurs BG (1994) A functional anatomical study of associative learning in humans. Proc Natl Acad Sci USA 91:8122–8126
3. Whalen PJ, Kapp BS, Pascoe JP (1994) Neuronal activity within the nucleus basalis and conditioned neocortical electroencephalographic activation. J Neurosci 14(3 Pt 2):1623–1633
4. Morgan MA, LeDoux JE (1995) Differential contribution of dorsal and ventral medial prefrontal cortex to the acquisition and extinction of conditioned fear in rats. Behav Neurosci 109(4):681–688
5. LeDoux J (1996) Emotional networks and motor control: a fearful view. Prog Brain Res 107:437–446
6. Büchel C, Morris J, Dolan RJ, Friston KJ (1998) Brain systems mediating aversive conditioning: an event-related fMRI study. Neuron 20(5):947–957
7. LaBar KS, Gatenby JC, Gore JC, LeDoux JE, Phelps EA (1998) Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. Neuron 20:937–945
8. Fischer H, Andersson JL, Furmark T, Fredrikson M (2000) Fear conditioning and brain activity: a positron emission tomography study in humans. Behav Neurosci 114:671–680
9. Davis M, Whalen PJ (2001) The amygdala: vigilance and emotion. Mol Psychiatry 6(1):13–34
10. Gottfried JA, O'Doherty J, Dolan RJ (2002) Appetitive and aversive olfactory learning in humans studied using event-related functional magnetic resonance imaging. J Neurosci 22(24):10829–10837
11. Herry C, Garcia R (2002) Prefrontal cortex long-term potentiation, but not long-term depression, is associated with the maintenance of extinction of learned fear in mice. J Neurosci 22(2):577–583
12. Dunsmoor JE, Bandettini PA, Knight DC (2008) Neural correlates of unconditioned response diminution during Pavlovian conditioning. Neuroimage 40:811–817

INTERNATIONAL
JOURNAL OF COMPARATIVE PSYCHOLOGY

Title:
Operant/Classical Conditioning: Comparisons, Intersections and Interactions The 2014 Winter Conference on Animal Learning and Behavior Focus and Research Seminar Sessions

Journal Issue:
International Journal of Comparative Psychology, 27(4)

Classic Conditioning

Author:
Weiss, Stanley Jerome, American University
Rosales-Ruiz, Jesús, University of North Texas

Publication Date:
2014

Publication Info:
International Journal of Comparative Psychology

Permalink:
http://escholarship.org/uc/item/4c46c9gg

Acknowledgements:
We would like to acknowledge the contribution made to this issue by those who insightfully and constructively reviewed the articles. The reviewers are, in alphabetical order: Björn Brembs, José E. Burgos, John Donahoe, Philip Hineline, Jeremie Jozefowiez, David Kearns, Peter Killeen, James MacDonall, Allen Neuringer, Bruce Overmier, David Palmer, Alan Silverberg and John Staddon.

Author Bio:
Professor of Experimental Psychology Emeritus

Associate ProfessorUniversity of North Texas

Keywords:
Operant Conditioning, Classical Conditioning, Conditioning and Genes, Law of Effect, Premack, Incentive Motivatio, Choice, Pavlovian-to-instrumental Transfer, Transfer of Control, Winter Conference on Animal Learning and Behavior

Local Identifier:
uclapsych_ijcp_24902

Abstract:
The Keynote Speaker at Winter Conference on Animal Learning and Behavior (WCALB) 2014 was Dr. Björn Brembs whose address was titled, Pavlovian and Skinnerian Processes Are Genetically

Separable. The essence of the address, that describes the research on which Dr. Brembs based this conclusion, is described below. Articles in this issue representing the related Focus Session include: The Many Faces of Pavlovian Conditioning by Dr. Jozefowiez, Pavlov + Skinner = Premack by Dr. Killeen, Evocation of Behavioral Change by the Reinforcer is the Critical Event in Both the Classical and Operant Procedures by Dr. Donahoe, On Choice and the Law of Effect by Dr. Staddon, Response-Outcome versus Outcome-Response Associations in Pavlovian-to-Instrumental Transfer: Effects of Instrumental Training Context by Gilroy, Everett and Delamater, and The Instrumentally-Derived Incentive-Motivational Function by Dr. Weiss. As a whole, they attempt to increase our contact with, and get at the essence of, what is actually happening with these operant and classical contingencies in the laboratory and nature. The Research Seminar Session revealed the current tendency for explanations of behavior to be reduced to physiology, neuroscience, and genetics. However, anti-reductionists saw shortcomings in this approach. They recommended an interconnected holistic approach which shifts the focus away from the structure of discrete behaviors and toward examining the environment in which the behavior occurs and the consequences produced. The distinction between structural and functional analysis points to a difficulty of integrating facts about behavior with other levels of analysis that requires our attention.

INTERNATIONAL
JOURNAL OF COMPARATIVE PSYCHOLOGY

2014, 27(4), 515-525
Stanley J. Weiss &
Jesús Rosales-Ruiz, Editors
Special Issue

Introduction to the Special Issue on Operant/Classical Conditioning: Comparisons, Intersections and Interactions

The 2014 Winter Conference on Animal Learning and Behavior Focus and Research Seminar Sessions

Classical Conditioning

Stanley J. Weiss
American University, USA

Jesús Rosales-Ruiz
University of North Texas, USA

The Keynote Speaker at Winter Conference on Animal Learning and Behavior (WCALB) 2014 was Dr. Björn Brembs whose address was titled, *Pavlovian and Skinnerian Processes Are Genetically Separable.* The essence of the address, that describes the research on which Dr. Brembs based this conclusion, is described below. Articles in this issue representing the related Focus Session include: *The Many Faces of Pavlovian Conditioning* by Dr. Jozefowiez, *Pavlov + Skinner = Premack* by Dr. Killeen, *Evocation of Behavioral Change by the Reinforcer is the Critical Event in Both the Classical and Operant Procedures* by Dr. Donahoe, *On Choice and the Law of Effect* by Dr. Staddon, *Response-Outcome versus Outcome-Response Associations in Pavlovian-to-Instrumental Transfer: Effects of Instrumental Training Context* by Gilroy, Everett and Delamater, and *The Instrumentally-Derived Incentive-Motivational Function* by Dr. Weiss. As a whole, they attempt to increase our contact with, and get at the essence of, what is actually happening with these operant and classical contingencies in the laboratory and nature. The Research Seminar Session revealed the current tendency for explanations of behavior to be reduced to physiology, neuroscience, and genetics. However, anti-reductionists saw shortcomings in this approach. They recommended an interconnected holistic approach which shifts the focus away from the structure of discrete behaviors and toward examining the environment in which the behavior occurs and the consequences produced. The distinction between structural and functional analysis points to a difficulty of integrating facts about behavior with other levels of analysis that requires our attention.

Dr. Björn Brembs Professor of Neurogenetics at the Institute of Zoology at Universität Regensburg was the Keynote Speaker at the 2014 Winter Conference on Animal Learning & Behavior. The title of his address was *Pavlovian and Skinnerian Processes are Genetically Separable.* This set the stage for the Conference theme and the related Focus Session concerned with *Operant/Classical Conditioning: Comparisons, Intersections and Interactions.* Articles in this issue of the *International Journal of Comparative Psychology* have been developed from that Focus Session. The impressive Keynote Address Dr.

Please send correspondence to Dr. Staley Weiss, American University, Washington, USA and Jesús Rosales-Ruiz, (Email: sweiss@american.edu and jesus.rosales-ruiz@unt.edu, respectively.

Brembs delivered is sketched below. For additional information, see Brembs and Plendl (2008), Brembs (2009, 2011) and Colomb and Brembs (2010). The slides from Dr. Brembs' presentation can be found at http://www.slideshare.net/brembs/wcalb14.

In his introduction, Brembs noted that Skinner realized in the 1930s that operant conditioning was likely composed of several processes (Konorski & Miller, 1937a, 1937b; Skinner, 1935, 1937). Brembs went on to claim that his research demonstrates that Skinner was right. He described an operant conditioning situation where a food-deprived rat was reinforced with a food pellet for every lever press (a fixed ratio or FR1), and went on to explain that, although probably there are two processes operating simultaneously here, this paradigm makes it difficult to reveal them directly. Instead, one needs to use other experiments, such as those he designed with tethered flies as subjects, organized analogously to the rats' FR1 situation, wherein he could precisely measure the direction in which they were turning.

Brembs explained that these experiments used colors instead of a lever: One color (say, blue) was associated with an aversive heat-beam – much like an unpressed lever is associated with no food. In contrast, another color (e.g., green) was associated with the heat beam being switched off when the stationary flying fly attempted a turn in the proper direction – much like a depressed lever produces food.

From the way he views these paradigms, in both experiments the stimuli were under full control of the animal. When the rat presses the lever or the fly generates a particular turning maneuver, the situation changes. For the rat, food is delivered. For the fly, heat is switched off. In both cases, the animal's behavior has produced a preferred situation. If the rat stops pressing and engages in some other behavior, food ceases to be delivered. If the fly stops attempting to turn to the right and instead attempts a left turn (coloring its environment blue), the heat is switched on.

Brembs claimed that if the rat case parallels the fly case such that:

> (a) the fly turning right under green light and shutting the heat off is analogous to the rat pressing the lever down and producing food, *and*
> (b) the fly turning left under blue light and turning the heat on is analogous to the rat not pressing the lever (lever up) leading to no food,

then two things happen to rats in a Skinner box. The rat learns that the lever means food (just as the fly learns that green means heat) and that lever pressing produces food. The topography of behavior (i.e., left or right paw, nose, tail, etc.) is irrelevant as long as the lever is pressed and the microswitch is closed. In this stage, (1) the behavior is still flexible, (2) any modification of the neural circuits controlling the behavior are actively suppressed and, (3) learning about the environment is dominant – what Brembs calls *world-learning*. He reports that the biochemical processes and associated molecules involved in this type of learning are those discovered in classical conditioning, e.g., adenylyl cyclases of type I, cyclic adenosine monophosphate (cAMP), and protein kinase A or PKA (Brembs & Plendl, 2008).

Brembs went on to explain that this memory of the environment is independent of the behavior that controlled the stimuli. That is, the animals can use a different, even orthogonal, behavior to press the lever (or change the color) and still keep their performance up, compared to controls (Brembs, 2009). In the case of the flies, he can test this hypothesis in a way that is impossible in the rat because if the lever is removed, there is nothing for the rat to press - if one wants to test if the behavior has been conditioned independently of the lever. In comparison, in his paradigm with flies, turning direction preference can be tested after training, with the colors turned off. At this stage, using this paradigm, the flies show no preference (Brembs, 2009).

The second thing that happens, Brembs elaborated, is that, as conditioning continues, the suppression or inhibition that keeps the behavior flexible is overcome – making the behavior more stereotyped. This means that now the behavior is not that flexible any more so a different behavior is less effective in controlling the stimuli. In the case of the flies, when Brembs removed the colors after this extended training, now they showed a preference for the previously unpunished turning direction. With this stage of conditioning having been termed *habit formation* in vertebrates, and because the fly situation appeared quite similar, Brembs co-opted this term for invertebrates as well.

In rats, after habit formation, selective devaluation of the reinforcer no longer leads to a reduction in lever pressing rate, demonstrating the behavior's inflexibility (Balleine & Dickinson, 1998; Balleine & O'Doherty, 2010; Costa, 2007, 2011; Graybiel, 2005; Yin & Knowlton, 2006). At this stage, the biochemistry underlying the recruited neuronal circuitry is completely different. However, if the type I cyclases function is blocked, then the flies can continue to learn well. Protein kinase C (PKC), instead of PKA, is important with PKC manipulations having no effect on world-learning. This type of learning modifies the behavior of the animal itself and Brembs therefore termed it *self-learning* (Colomb & Brembs, 2010).

Brembs went on to explain that, in flies, it can be showed that it is the inhibition of world-learning that slows down self-learning. Compared to learning with colors, the same experiment without colors leads to a PKC-dependent preference for turning direction requiring only half the training sessions (Brembs & Plendl, 2008).

Besides PKC, Brembs showed that the gene *FoxP* is also necessary for self-learning. He explained that this is important because the genes *FOXP1* and *FOXP2* in humans (the closest relatives to *FoxP* in flies) are associated specifically with language acquisition and other motor skills. In fact, he asserts that the *FoxP* evidence directly contradicts Chomsky's statement that language acquisition and operant learning share only superficial similarities (Mendoza et al., 2014).

Finally, Brembs described how in the marine snail *Aplysia*, he and his colleagues have isolated the neurons involved not only in learning which behavior is rewarded/punished, but also the ones controlling when the behavior is supposed to be initiated (Brembs, 2003; Brembs, Lorenzetti, Reyes, Baxter, & Byrne, 2002; Nargeot, Le Bon-Jego, & Simmers, 2009; Nargeot & Simmers, 2010, 2012; Sieling, Bédécarrats, Simmers, Prinz, & Nargeot, 2014). This research shows how operant learning modifies both the timing and the kind of behavior that will be emitted. Moreover, it provides mechanistic insight in how behavioral variability is first

generated on a neuronal level and then restricted due to reinforcement. Brembs contends that these groundbreaking findings provide a roadmap for future research on the neurobiological processes underlying all forms of operant learning in vertebrates as well as in invertebrates.

Special Issue Article Overview

The following articles in this issue, by highly regarded behavioral scientists concerned with conditioning and learning, contribute to a more complete appreciation of the intersections, interactions, and commonalities, as well as the differences between operant and classical conditioning processes. As a whole, they serve to increase our contact with, and get at the essence of, what is actually happening with these contingencies in the laboratory as well as in nature. Each of the articles speaks for itself, but there are some themes, objectives, consolidations and concerns represented therein that will be briefly touched on below.

Jozefowiez in *The Many Faces of Pavlovian Conditioning* shows how included under the term "Pavlovian conditioning" are: a procedure, the learning resulting from that procedure plus the process meant to explain that learning. He is concerned that this interferes with our adequately understanding these various phenomena. In fact he shows even if one focuses at only the *process* there is behavioral as well as neural evidence supporting that it is "extremely unlikely that a single Pavlovian conditioning process is responsible for learning in all procedures classified as Pavlovian conditioning." Jozefowiez describes how the same concerns also apply to the operant conditioning situation.

Pavlovian conditioning is the name of a procedure in which a CS is paired with a US and Jozefowiez describes a wide range of preparations exemplifying this procedure. They include: Pavlov's salivary conditioning procedure in dogs, the nictitating membrane response conditioning in rabbits, fear conditioning in rats, autoshaping in pigeons and taste aversion learning. In each case, a CS is paired with a US and the CS comes to elicit a response related to the US.

Weiss' *Instrumentally Derived Incentive Motivational Functions* (appetitive and aversive) may add "faces" to Pavlovian conditioning as a procedure as well as a process. He proposes that these functions support the contention that environmental conditions being differentially correlated with operant-behavior-produced reinforcement could be responsible for most classical conditioning and resulting incentive-motive states." There was no traditional Pavlovian CS – US pairing in any of his training conditions. This can be considered an example of how learning principles derived from the laboratory can be extended to help us better understand how they are likely operating in nature. These "states" are clearly learned, anticipate the future and energize behavior.

Donahoe *is* seeking the "essentials." He proposes "... that the critical reinforcing event in both the classical and operant procedures is more closely correlated with the evocation of a *change in behavior* induced by the reinforcing stimulus than with the presentation of that stimulus itself." He goes on to describe evidence at the behavioral, as well as the neurological level, that support this consolidation. This moment-to-moment

conception of the conditioning process represents an insightful integration. Weiss proposes that most classical conditioning occurs outside the laboratory through the reinforcement differences that come to exist between environmental situations as the underlying operant contingencies take hold. Might a more molar model be necessary for the "classical conditioning" produced in this manner?

In *Pavlov + Skinner = Premack,* Killeen makes Premack's principle central to all reinforcement. Essentially, Premack translated *reinforcement* from what the operant produced *per se* to the behavior the subject engaged in, after the operant created the appropriate opportunity. According to this Premack, when one's preferred behaviors are ranked from 1 to n, a behavior is reinforced by the opportunity to engage in a lower ranked one and punished by producing conditions that require one to engage in a higher ranked behavior.

Premack originally tested his principle in rather simple laboratory situations wherein he showed, for example, that children with an initial preference for pin-ball playing over gum drops, would eat gum drops for the opportunity to play pin-ball, and vice-versa (Premack, 1959). Although confirming the process, studies of this type probably also limited and constrained its application.

Killeen takes Premack out of the laboratory and uses a hierarchy of natural action patterns, adapted from Timberlake and Lucas (1989), that applies Premack's principle to complex behavioral repertoires of animals in nature. It is Killeen's thesis "...that operations that move an animal into a particular subsystem or mode [e.g., predation, defense, escape] are *motivational.* When in that subsystem, the ability to engage in an action ... will reinforce actions that lie above it ... and be reinforced by actions that lie below it in that column."

He goes on to describe how through selective breeding the English Pointer is trapped in one link of an action chain and perpetually unable to consummate the nominal goal of that action – and is concerned that our laboratory animals have regularly been stuck at local minima. Killeen clearly helps us appreciate how (a) operant conditioning, (b) naturally occurring sequences of behavior related to particular situations and conditions plus, (c) Pavlovian conditioning can be effectively integrated for a more complete understanding of-going behavior.

Staddon, in his *On Choice and the Law of Effect* (LOE), presents a simple dynamic analysis of the molar matching law and related research. He begins with cumulative records which show behavior of individulal subjects in real time – rather than thousands of aggregated responses as most "matching" experiments do. His simple law-of-effect (LOE) model easily reproduces the post-reinforcement "burst" of responses that reliably occur when animals are first exposed to intermittent reinforcement. He shows how much can be explained just by this elementary, non-temporal process.

He goes on to explain how concurrent variable-interval schedules tend to produce alternation because of the strong negative feedback that VI schedules provide – the longer one choice is neglected, the higher its payoff probability becomes. It is perhaps not surprising that a changeover delay is usually needed to get the

matching relation: "The precise correspondence between relative frequency of responding and relative frequency of reinforcement broke down when the COD was omitted." (Herrnstein, 1961, p. 271).

Staddon points out that concurrent random-ratio schedules have no such feedback. Data from random-ratio schedules suggest a law-of-effect-based cumulative-effects model which assumes: (a) that reinforcement-linked response tendencies compete; (b) that the highest-strength response is the one that occurs; (c) other responses are "silent," their strength unchanged until they occur; and (d) response strengths depend on relatively long-term reinforcement history. He shows that LOE models make several interesting predictions where matching either fails or does not apply.

Staddon's discussion of VI schedules in the matching research led the author (SJW) to recall that some years ago he ran a few undergraduates on a VI schedule for several 30-min sessions as a baseline for an anticipated subsequent manipulation. When interviewed after these sessions, every student said, to the effect, "Well Professor Weiss, you stumped me. A few times I thought I figured out the sequence, like two fast and three spaced responses, but ultimately I couldn't." That humans are accustomed to solving problems could severely limit the application of findings based on paradigms without a "logical solution".

Gilroy, Everett and Delamater, in their *Response-Outcome versus Outcome-Response Associations in Pavlovian-to-Instrumental Transfer: Effects of Instrumental Training Context*, report that training two instrumental Response-Outcome (R-O) relations in distinct contexts disrupted the ability of those responses to be modulated by Pavlovian stimuli in a sophisticated multiple response, outcome and context Pavlovian-to-Instrumental (PIT) paradigm. They interpret this to support the view that instrumental learning results in the development of highly specific Response-Outcome (R-O) associations that are used in the backward direction during a specific PIT test.

Gilroy et al. go on to discuss their results in the context of the potential associations that could be learned (context-outcome, response-outcome, and outcome-response) during instrumental learning and caution that although "... we have emphasized associative mechanisms at the intersection of Pavlovian and instrumental learning that would permit for an interaction at the levels of learning and performance, it is important to realize that there very well may be important differences, as well, in the underlying associative circuitries of Pavlovian and instrumental learning that may prevent interactions from taking place in some circumstances." (It should be noted that PIT is the current manner of describing what has traditionally been referred to as the transfer-of-control (TOC) paradigm. See Weiss' article in this issue for a functional analysis of the single response, outcome and context TOC design.)

Research Seminar Session: Conditioning, Learning, Behavior and Levels of Explanation

The status of behavioral research as a scientific subject matter is well established. Yet, even though behavioral analysis is a prominent feature of various disciplines (e.g., behavioral genetics, behavioral

neuroscience, behavioral biology and behavioral pharmacology), it too often continues to be regarded as relevant only because of what it tells us about other processes considered, more important such as physiology and brain function. Skinner (1938) wrote, "Facts about behavior are not treated in their own right, but are regarded as something to be explained or even explained away by the prior facts of the nervous system." (p. 4). This reductionistic tradition continues to be favored today. West and King (2001) summarized it well when they said, "... the study of behavior seems to be something someone does to get to somewhere else. Behavior affords a gateway to physiology, neuroscience, and genetics." (p. 587).

This issue of reductionism was raised at the WCALB 2014 Research Seminar Session. In many ways, the points discussed paralleled current debate in the scientific community. For some participants, it was already all said and done that behavioral phenomena could be explained completely by physiology, neuroscience, and genetics and by reducing behavior to neural networks. The behavioral explanation was questioned, since it was assumed that the more complete explanation of behavior must be found inside the black box. This position was opposed by some who asserted instead that an explanation outside the black box at the behavioral level was valid and the way to go.

However, those who have searched for explanations at the behavioral level have been accused of focusing only on behavior and ignoring the role of physiology, neuroscience, and genetics. This is unfortunate, because without those behavioral facts, physiology, neuroscience, and genetics would not have much to study. As Skinner said, "Valid facts about behavior are not invalidated by discoveries concerning the nervous system, nor are facts about the nervous system invalidated by facts about behavior. Both sets of facts are part of the same enterprise, and I have always looked forward to the time when neurology would fill in the temporal and spatial gaps which are inevitable in a behavioral analysis." (1988, p. 128). Thus, physiology, neuroscience and genetics supplement a behavioral analysis at the level of the whole organism interacting with the environment. From this perspective, facts about the organism (i.e., physiology, brain and genes) are considered part of the conditions necessary for the occurrence of behavior rather than explanations of behavior on their own. The explanation lies in the environmental contingencies, but facts from neuroscience and physiology help fill in the physical details that connect behavior to biology.

One alternative to reductionism discussed at WCALB 2014 was to acknowledge that there are different causes of behavior as proposed by Aristotle (triggers, functions, substrate and formal representation), all valid in their own right. In this view, the understanding of any phenomenon requires the understanding of all of its causes and how they interconnect. This position is in keeping with the modern anti-reductionism stance. For instance, in recent years biological scientists have moved away from the idea that biological systems can be fully explained by molecular biology and have instead favored some sort of holism or interconnected system (Van Regenmortel, 2004).

In the study of development, West and King (2001) have also recommended a systems approach that integrates the molecular and the non-molecular. Similarly, in recognition that drug-taking cannot be reduced to drug actions, those who study addictions have also called for a systems approach (see Kalant, 2009; and commentaries). A systems approach, while not easy, has the advantage of letting different disciplines pursue

explanations at different levels as parts of a bigger picture. Complete understanding of behavior then comes from the coordination of all of these explanations (see, Killeen, 2001). An integrative science of behavior depends, of course, on the integrity of the levels. Errors at one level would necessarily produce errors in the explanatory system (Ruff, 2003).

West and King (2001) caution that the linkage between levels is compromised when the behavioral phenomena linked are devoid of important ecological and social variables relevant to the occurrence of the behavior in question. For example, they found that despite an elegant and thorough analysis of song-related neural structures, this analysis alone could not predict copulatory success. Instead, they discussed that they "... have taken a more direct approach to understanding the task of song development by focusing on its goal, successful courtship." (p. 597). As researchers move toward looking at behavior in its ecology, the focus is shifting away from the structure of discrete behaviors and instead toward examining the environment in which the behavior occurs and the consequences produced.

The distinction between structural and functional analysis points to one difficulty of integrating facts about behavior with other levels of analysis. Before behavior can be linked to other levels of observation or even reduced to one of those levels, there needs to be agreement about what is being reduced or integrated. Physiology, neuroscience and genetics have traditionally relied on a structural analysis of behavior rather than a functional one. A bird's song no doubt has a biological structure that can be mapped to other structures and functions of the nervous system. However, focusing only on structure takes away from the environmental variables that are part of the behavior. For example, a bird song could be one behavior when it is directed toward females, yet the same song could be a different behavior when directed toward males or when emitted in the absence of conspecifics. Although the song remains the same, the interaction with the environment produces different consequences. From a behavior analytic point of view, explanations are concerned with stating functional relations between behavior change and changes in the antecedent and consequent environment. These functional relations between behavior and environment create analytic units that make possible the study of behavior.

At WCALB the issue of structural and functional analysis of behavior was also discussed but at a more basic level: The level of the stimulus and response. Participants debated whether it is possible to talk about a stimulus without reference to a response or a response without reference to a stimulus. This question was new to the neuroscientists in the discussion, who were more familiar with defining stimuli and responses with respect to their structural and physical characteristics only. From this approach, the properties of the response can be defined completely by the field of biology and the properties of the stimulus by the field of physics. For the behavior scientists, however, the stimulus and response are two factors that interact to make up their subject matter: behavior. The response by itself does not constitute a behavioral event, even though it can be defined as a more or less fixed pattern of neural, muscular and glandular events all the way down to its molecules. The stimulus must also enter into the definition of behavior and it is this connection between stimulus and response that gives rise to behavioral phenomena. While behavior scientists begin with the physical description of the stimulus and response, just as biologists and physicists do, these descriptions are insufficient to characterize the interactions that behavior scientists study.

Another question raised during the discussion was what to call a stimulus that is not connected to a response or a response that is not connected to a stimulus. It was suggested to call a response not connected with a stimulus an *action*. This led to the next question, is there such a thing as a response without a stimulus? For the determinists in the room, that could not be the case. If there was a response, they would search for the stimulus. This was considered an error by the neuroscientists because it would prevent the search for internal mechanisms of action. However, the determinists objected to dissecting the stimulus from the response. Behavioral phenomena are continuous. A response leads to a stimulus which in turn leads to another response and so on, as Killeen's article in this issue describes. This continuous nature of behavior is central to the analysis of behavior.

The discussions at WCALB were very productive and stimulating. They reflected the scientific scene in general, in that an integration of the various disciplines is a continuous and ongoing challenge. It became clear that in order to have an integrative science of behavior more discussions are needed regarding the basic assumptions of the various disciplines concerned with explaining behavior. What is a stimulus, what is a response, what is the nature of the interactions that the different disciplines consider important? The mingling of behavior scientists and neuroscientists at WCALB was a good step toward addressing these questions. The meeting ended with a reminder that "criteria" is plural. We all need to be mindful of the other's perspective.

The video of the Research Seminar Session can be found at:

https://www.youtube.com/watch?v=usCYJy9Pl9Q&list=UUIq9EAG5h9yc25rYSjywNMw.

We want to thank Scott Cohn for recording this video.

References

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology, 37*, 407-419.

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology, 35*, 48–69. doi:10.1038/npp.2009.131

Brembs, B. (2003). Operant reward learning in aplysia. *Current Directions in Psychological Science, 12*, 218–221. doi:10.1046/j.0963-7214.2003.01265.x

Brembs, B. (2009). Mushroom bodies regulate habit formation in Drosophila. *Current Biology, 19*, 1351–1355. doi:10.1016/j.cub.2009.06.014

Brembs, B. (2011). Spontaneous decisions and operant conditioning in fruit flies. *Behavioural Processes, 87*, 157–164. doi:10.1016/j.beproc.2011.02.005

Brembs, B., Lorenzetti, F. D., Reyes, F. D., Baxter, D. A, & Byrne, J. H. (2002). Operant reward learning in Aplysia: neuronal correlates and mechanisms. *Science*, 296, 1706–1709. doi:10.1126/science.1069434

Brembs, B., & Plendl, W. (2008). Double dissociation of pkc and ac manipulations on operant and classical learning in drosophila. *Current Biology, 18*, 1168–1171. doi:10.1016/j.cub.2008.07.041

Colomb, J., & Brembs, B. (2010). The biology of psychology: Simple conditioning? *Communicative & Integrative Biology, 3*, 142–145. doi:10.4161/cib.3.2.10334

Costa, R. M. (2007). Plastic corticostriatal circuits for action learning: what's dopamine got to do with it? *Annals of the New York Academy of Sciences, 1104*, 172–91. doi:10.1196/annals.1390.015

Costa, R. M. (2011). A selectionist account of de novo action learning. *Current Opinion in Neurobiology, 21*, 579–86. doi:10.1016/j.conb.2011.05.004

Graybiel, A. (2005). The basal ganglia: learning new tricks and loving it. *Current Opinion in Neurobiology, 15*, 638–644. doi:10.1016/j.conb.2005.10.006

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior, 4*, 267-272.

Kalant, H. (2009). What neurobiology cannot tell us about addiction. *Addiction, 105*, 780–789

Killeen, P. (2001). The four causes of behavior. *Current Directions in Psychological Science, 10*, 136-140.

Konorski, J., & Miller, S. (1937a). Further remarks on two types of conditioned reflex. *Journal of General Psychology, 17*, 405–407.

Konorski, J., & Miller, S. (1937b). On two types of conditioned reflex. *Journal of General Psychology, 16*, 264–272.

Mendoza, E., Colomb, J., Rybak, J., Pflüger, H.-J., Zars, T., Scharff, C., & Brembs, B. (2014). Drosophila FoxP mutants are deficient in operant self-learning. *PloS One*, *9*, e100648. doi:10.1371/journal.pone.0100648

Nargeot, R., Le Bon-Jego, M., & Simmers, J. (2009). Cellular and network mechanisms of operant learning-induced compulsive behavior in Aplysia. *Current Biology, 19*, 975–84. doi:10.1016/j.cub.2009.05.030

Nargeot, R., & Simmers, J. (2010). Neural mechanisms of operant conditioning and learning-induced behavioral plasticity in Aplysia. *Cellular and Molecular Life Sciences, 68*, 803-816. doi:10.1007/s00018-010-0570-9

Nargeot, R., & Simmers, J. (2012). Functional organization and adaptability of a decision-making network in aplysia. *Frontiers in Neuroscience, 6*, 113. doi:10.3389/fnins.2012.00113

Premack, D. (1959). Toward empirical behavior laws: I. Positive reinforcement. *Psychological Review, 66*, 219-233.

Ruff, R. M. (2003). A friendly critique of neuropsychology: facing the challenges of our future. *Archives of Clinical Neuropsychology, 18*, 847–864

Sieling, F., Bédécarrats, A., Simmers, J., Prinz, A. A., & Nargeot, R. (2014). Differential roles of nonsynaptic and synaptic plasticity in operant reward learning-induced compulsive behavior. *Current Biology, 24*, 941–950. doi:10.1016/j.cub.2014.03.004

Skinner, B. F. (1935). Two types of conditioned reflex and a pseudo type. *Journal of General Psychology, 12*, 66–77.

Skinner, B. F. (1937). Two types of conditioned reflex: A reply to Konorski and Miller. *Journal of General Psychology, 16*, 272–279.

Skinner (1938). *The behavior of organisms*. New York, NY: Appleton-Century-Crofts.

Skinner (1988). Reply to Marc N. Richelle. In A. C. Catania & S. Harnad (Eds.), *The selection of behavior. The operant behaviorism of B. F. Skinner: Comments and consequences* (p. 128). New York, NY: Cambridge University Press.

Timberlake, W., & Lucas, G. A. (1989). Behavior systems and learning: From misbehavior to general principles. In S. B. Klein & R. R. Mowrer (Eds.), *Contemporary learning theories: Instrumental conditioning theory and the impact of constraints on learning* (pp. 237-275). Hillsdale, NJ: Erlbaum.

Van Regenmortel, M. H. V. (2004). Reductionism and complexity in molecular biology. *EMBO Rep*orts, *5*, 1016-1020.

West, M. J., & King, A. P. (2001). Science lies its way to the truth ... really. In E. M. Blass (Ed.), *Developmental Psychobiology: Handbook of Behavioral Neurology,* (Vol. 13, pp. 587-614). New York, NY: Kluwer Academic/Plenum Publishers

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews. Neuroscience, 7*, 464–476. doi:10.1038/nrn1919

Financial Support: Preparation of this work was supported in part by NIDA Award R01-DA008651. It is solely the authors' responsibility and doesn't necessarily represent official views of NIDA/NIH.

Conflict of Interest: The authors declare no conflict of interest.

Behaviorism

Learning Theories: Behaviorism

Kevin R Clark, EdD, R.T.(R)(QM)

In its simplest form, learning is defined as gaining knowledge through study, teaching, instruction, or experience.[1] Interestingly, learning is described and viewed differently by theorists, researchers, and practitioners who have spent time investigating and experimenting in the educational psychology field.[1,2] The differences in how educational theorists believe individuals acquire, retain, and recall knowledge resulted in the development of multiple learning theories.[1-3] Based on the context of the theorists' work and other factors at the time of investigation, these theories explain how learning occurs, what internal or external factors influence learning, how memory affects learning, and how transfer of knowledge occurs.[1-3] In addition, the roles of the instructors and learners are described according to each theory of learning. A basic understanding of the various learning theories is essential for educators who strive to lead a classroom that is conducive to learning and success.

The ideas of behaviorism date back to the late 19th and early 20th centuries when John Watson, an American psychologist, believed the general public would accept and recognize the new philosophy of psychology as a true science only if it involved processes of objective observation and scientific measurement.[1] This notion of detailed observation and measurement became central to the work of behaviorists.[1]

Behaviorism emphasizes that learning occurs when an individual responds favorably to some type of external stimuli.[1-4] Behaviorism sometimes is referred to as the *stimulus-response theory*.[1] For example, when presented with a math flashcard showing the equation 6×8, the learner responds with the answer 48. The equation is the stimulus, and the answer is the associated response.[2] Essential elements with behaviorism include the stimulus, the response, and the association between these 2 elements.[2] Of particular importance is how the association between the stimulus and the response is made, strengthened, and maintained.[2]

Behaviorists define learning as nothing more than the acquisition of new behaviors. Behaviorists do not emphasize thinking or other mental activities as a part of the learning process because such variables are not observable behaviors.[1-4] Although the behaviorism theory discounts any mental activity, other educational theorists considered these processes to be a vital part of learning and cognition, which resulted in the development of other theories of learning.[1,4] Behaviorists do not address memory and how new behaviors or changes in behaviors are stored or recalled for future use.[2] Behaviorists refer to this type of learning, where a reaction is made to a particular stimulus, as *conditioning*.[1] Two main types of conditioning include Pavlov's classical conditioning and Skinner's operant conditioning.

Classical Conditioning

Ivan Pavlov, a Russian physiologist, noticed that dogs salivated every time they ate or saw food and believed

he could condition the dogs to salivate at the sound of a bell.[1] Initially, Pavlov sounded a bell at the time food was presented to the dogs and repeated this process frequently.[1] Eventually, the sound of the bell became an indication to the dogs that food was about to be presented, and they responded by salivating at the sound of the bell regardless of whether food was presented.[1] This type of reinforcement of a natural reflex or some involuntary behavior that occurs as a response to a particular stimulus is called *classical conditioning*.[1] Pavlov was able to condition the dogs to salivate in response to the sound of the bell.

Pavlov identified 4 stages of classical conditioning: acquisition, extinction, generalization, and discrimination.[1] The acquisition stage is the initial learning of the conditioned response (the dogs salivating at the sound of the bell).[1] Pavlov believed the conditioned response would not remain indefinitely, so he used the term *extinction* to describe the disappearance of a conditioned response.[1] Pavlov demonstrated extinction by repeatedly sounding the bell without presenting food to the dogs.[1] The final 2 stages, *generalization* and *discrimination*, are opposites and explain how behaviorists believe knowledge is transferred within learners.[2] The generalization stage implies that a conditioned response might occur with similar stimuli without further training (the dogs salivating at the sound of something similar to a bell).[1] In contrast, the discrimination stage indicates that a conditioned response might occur with 1 stimulus but not with another (the dogs not salivating at the sound of something similar to a bell).[1]

Operant Conditioning

BF Skinner, a psychologist working in the United States in the 1930s, established the theory of operant conditioning: a process of reinforcing a voluntary behavior by rewarding it.[1,3] Studying the behaviors of rats, Skinner used a device (now called a *Skinner box*) that contained a lever.[1] Whenever the rats pressed the lever (an action Skinner considered normal, random, and voluntary), a pellet of food was presented.[1] As the food rewards continued during the repetition of the action, the rats learned that they had to press the lever to be fed.[1] Skinner also used reinforcement techniques to teach pigeons to dance and to roll a ball down a mini bowling alley.[1] Skinner made generalizations about his findings with rats and pigeons to humans.[1] In addition, he noted that operant conditioning also worked in a negative way: stopping a behavior from occurring by punishing it.[1,3]

Reinforcement and Punishment

Key aspects of operant conditioning include reinforcement and punishment, both of which can be positive or negative. Reinforcement refers to anything that has the effect of strengthening a particular behavior for it to occur again.[1,3] Positive reinforcement is the addition of a rewarding stimulus to get the behavior to happen again (eg, rewarding learners for making a high grade on an exam in hopes they study harder for future assessments and score high again). Negative reinforcement is the removal of an unpleasant stimulus to get the behavior to continue (eg, students learning the rules to solve a particular problem so their instructor quits nagging them about the importance of it). The unpleasant behavior of the instructor's nagging is removed when students learn the rules, solve the problem correctly, and continue the action so the nagging does not return.

Conversely, punishment refers to anything that has an effect of lessening or discouraging a particular behavior so that it does not occur again.[1,3] Positive punishment is the addition of an unpleasant stimulus to get the behavior to stop; any type of disciplinary action is considered positive punishment. Negative punishment is the removal of a rewarding stimulus to get the behavior to stop (eg, not offering extra credit opportunities in hopes the behavior stops so that the learners can receive these beneficial opportunities in the future). Skinner maintained that rewards and punishments control most human behaviors.[1-3]

In addition to Watson, Pavlov, and Skinner, other theorists were associated with the behaviorist movement. The **Table** summarizes their contributions to the theory of behaviorism.

Implications in Teaching and Learning

Behaviorists believe learning begins when a cue or stimulus from the environment is presented, and the learner reacts to the stimulus with some type of response.[1-3] Those responses are reinforced or punished,

Table

Key Theorists and Their Contributions to Behaviorism[1]

Theorists	Contribution
Ivan Pavlov	Classical conditioning
Edward Thorndike	Connectionism (emphasized the role of experience in the strengthening and weakening of stimulus-response connections)
John Watson	Scientific objectivity; Law of frequency (the more frequent a stimulus and response occur in association with each other, the stronger the habit will become)
Edwin Guthrie	Contiguity (the same response to a stimulus most likely will occur over and over again during repeated exposures)
BF Skinner	Operant conditioning

and this process is repeated so that the responses become automatic.[3] Ultimately, the change in behavior indicates learning has occurred.[3] As revealed, behaviorism has little regard for mental processes or understanding and, therefore, does not prepare learners for problem-solving or critical-thinking skills.[1-3]

The instructor plays a dominant role in behaviorism by leading the learning environment, using positive and negative reinforcement to shape learners' behaviors, and presenting the content.[1] With behaviorism, learners are described as passive individuals who voluntarily respond to external stimuli.[1] Other behaviorist implications in teaching and learning include[1]:

- creating procedures and expectations to manage the classroom
- using rewards as incentives for learners to work hard and behave
- using punishments (eg, loss of privileges or withholding of rewards) effectively and sparingly to change learners' behaviors

Critics of behaviorism argue that rewards can belittle or demean a learning experience and, therefore, should be used with caution.[1] Often, rewards can evoke feelings of unfairness or competition, and some learners might become distracted from the real issue involved in completing a task or learning new material.[1] Using a rewards system or giving 1 learner increased attention might have a detrimental effect on others in the class or cause them to feel isolated.[1] Not surprisingly, rewards do not always lead to higher-quality work; however, using a behaviorist approach, rewards can result in the reinforcement of appropriate classroom behaviors, which can create a more orderly classroom environment that is conducive to learning and success for all.[1]

Learning Activities

Classroom learning activities connected to the behaviorism theory include[1-3]:

- lecturing
- recalling facts
- defining and illustrating concepts
- applying explanations
- participating in rote learning (ie, memorization based on repetition)
- completing drill and practice exercises
- establishing classroom management policies
- using rewards and punishments

Implications in Medical Imaging Education

In medical imaging education, lecturing is a dominant approach to presenting information because of the complexity of the content. Considering time management issues and restrictions in higher education, lecturing affords instructors an opportunity to present a large amount of information to a large audience. Often, medical imaging students memorize some of the content presented and recall that knowledge during an exam. The role of repetition aids in the learning of new and challenging content. Medical imaging students benefit from drill and practice exercises when working with formulas, including the Inverse-Square Law, the milliampere-seconds–distance compensation formula,

and the grid conversion formula, as well as calculations involving skin dose. Medical imaging instructors benefit from using a behaviorist approach by implementing a classroom management plan to lead a classroom conducive to learning and success.

Conclusion

The theory of behaviorism can be illustrated by the adage, "practice makes perfect." Behaviorists see learning as an observable change in behavior as a result of experience and repetition. This stimulus-response theory makes no attempt to assess the mental processes necessary for learners to acquire, retain, and recall information. The change in behavior is simply achieved through a conditioning process using reinforcement and punishment. Even though little importance is placed on mental activity, concept formation, or understanding, there is a place for behaviorism in today's classrooms, especially in medical imaging education, in the areas of rote learning and classroom management.

Kevin R Clark, EdD, R.T.(R)(QM), is assistant professor and graduate coordinator for the School of Health Professions at The University of Texas MD Anderson Cancer Center in Houston. He serves on the Radiologic Technology *Editorial Review Board and can be reached at krclark@mdanderson.org.*

References

1. Pritchard A. Behaviourism and the beginnings of theory. In: *Ways of Learning – Learning Theories and Learning Styles in the Classroom.* 3rd ed. New York, NY: Routledge; 2014:6-17.
2. Ertmer PA, Newby TJ. Behaviorism, cognitivism, constructivism: comparing critical features from an instructional design perspective. *Perform Improv Q.* 2013;26(2):43-71. doi:10.1002/piq.21143.
3. Kelly J. Learning theories. The Peak Performance Center website. http://thepeakperformancecenter.com/educational-learning/learning/theories/. Published September 2012. Accessed June 10, 2017.
4. David L. Behaviorism. Learning Theories website. https://www.learning-theories.com/behaviorism.html. Published January 31, 2007. Accessed June 10, 2017.

Theoretical Behaviorism, Economic Theory, and Choice

John Staddon

Behaviorism

Psychology is not a single science. There are—alas—many psychologies. Behaviorism is one of them, and "theoretical" is one kind of behaviorism.[1] Economics also is far from unified. To understand the relation between psychology and economics, it might be best to begin with one psychology and one economics, rather than attempt a review of two highly various fields. I compare two ways to look at choice behavior: *theoretical behaviorism* and *prospect theory*, a popular approach to behavioral economics as it has been presented in Daniel Kahneman and Amos Tversky's seminal 1979 paper and later writings.

What is *choice*? Charles Darwin wrote copious notes, two months before his engagement to his cousin Emma Wedgwood, listing the pros and cons of marriage.[2] A sample:

> *Marry* [pros]:
> Children—(if it Please God)—Constant companion,
> (& friend in old age) who will feel interested in one,—

Correspondence may be addressed to John Staddon, Department of Psychology and Neuroscience, 242 Soc/Psych Building, Duke University, Durham, NC 27708; e-mail: jers@duke.edu.
Many thanks to Neal De Marchi and Marina Bianchi for the opportunity to participate in this conference.

1. For a review of the behaviorisms, see Staddon 2014a. For one attempt to apply theoretical behaviorism to the financial system, see Staddon 2012. An early effort to bring together psychology, biology, and economics is a collection of papers in Staddon 1980.

2. See www.darwinproject.ac.uk/darwins-notes-on-marriage.

History of Political Economy 48 (annual suppl.) DOI 10.1215/00182702-3619334

object to be beloved & played with.—better than a
dog anyhow.—Home, & someone to take care of
house—Charms of music & female chit-chat.—These
things good for one's health.—but *terrible loss of*
time . . .
Not Marry [cons]: Freedom to go where one liked—choice of Society
& little of it.—Conversation of clever men at clubs—
Not forced to visit relatives, & to bend in every
trifle.—to have the expense & anxiety of children—
perhaps quarrelling—Loss of time.—cannot read in
the Evenings—fatness & idleness . . .

There is much more, but you get the idea. This is choice indeed! No one would imagine that animals choose like this. Indeed, they do not, and for the most part, people do not either. So, I begin with the simplest possible choice situation. The two-armed bandit is popular not just in Las Vegas but among operant conditioners, those who study learned instrumental behavior. The subjects are usually animals—more easily, and ethically, subjected to experimental conditions than human beings and probably easier to understand. Pigeons and rats often seem to follow elementary economics. They are *rational*, in the sense that they maximize the rate of reward. For example, a hungry pigeon in a Skinner box randomly paid off on average for one in every ten pecks (termed a *random-ratio* schedule) on a response key on the left and on average for one in every five pecks for pecking on the right—such a pigeon will soon peck only on the right. It *maximizes*.

Indeed, optimality theory can explain many characteristics of the operant behavior of animals, of the ways they adapt to different kinds of *reinforcement schedules*. But animals also fail to act rationally, even in some strikingly simple situations.[3] The failures are more interesting than the successes. Sticking with ratio schedules, consider a *fixed*-ratio schedule that requires exactly one hundred pecks for each food reward. A pigeon will soon learn to get food this way. But he will always wait a second or two after each reward before beginning to peck, thus delaying food delivery unnecessarily. Fatigue? No; on a comparable *random* ratio, he will respond steadily, not waiting after food.

3. See, e.g., Staddon 2007 and references therein. For a comprehensive treatment, see Staddon 2016.

The reason he waits is that he has a built-in, automatic timing mechanism that responds to the minimum time between food deliveries enforced by the fixed (but not the random) ratio by pausing after each food. It is easy to show this by comparing two kinds of time-based schedules. Suppose you start a sixty-second timer after each food delivery and do not reward a peck until sixty seconds have passed (a fixed-interval schedule). As you might expect, once the animal has learned, he waits perhaps thirty seconds after each reward before beginning to peck. If his time sense were perfect, no variability, he would presumably wait exactly sixty seconds so as not to waste pecks. The pigeon's timing is a bit variable, so he usually starts early. Perfectly rational: do not waste pecks, but do not delay food unnecessarily. Now suppose you modify the procedure slightly by starting the sixty-second timer only after the first peck. The rational adaptation is simple: respond immediately after food, and then wait as before. But animals do not do this. If the interval duration is T seconds and the typical wait time is $T/2$, then on this response-initiated-delay schedule they will wait not $T/2$ but T seconds, delaying reward quite unnecessarily.

The message: optimality theory is of limited use in understanding what these animals are doing. Animals never, and humans almost never, are *explicit maximizers*, computing marginals and choosing accordingly, as in some interpretations of rationality. They adapt via built-in processes, like timing, that sometimes yield "rational" behavior and sometimes do not. Behavioral ecologists call them "rules of thumb" or (a familiar term to decision theorists) heuristics.

The time dimension adds complexity, and the way in which time is incorporated into learned behavior is still not perfectly understood. So, let us stick with random-ratio schedules, where time is irrelevant. A choice situation where there is no rational strategy or, to put it a little differently, all strategies are rational is choice between two *identical* random ratios. What do pigeons do and how might theoretical behaviorism explain it? The explanation follows the Darwinian consensus: instrumental learning is selection and variation. There is a repertoire of response tendencies—candidates. Candidates could be simple responses, heuristics, or what Gerd Gigerenzer calls an "adaptive toolbox" (Gigerenzer and Selten 2001). But only one response occurs at a time. In the example I am about to give, these candidates compete in a nonlinear, winner-take-all way, so that only the strongest tendency, the *active* response occurs. The other responses are *silent*—like recessive genes. The active response is either strengthened (if it is followed by reward) or diminished in strength (if there is no

reward). Notice that the variation-selection approach allows for individual differences. Different people will come to a given situation with different repertoires and different active responses.

Many dynamic processes have been proposed to describe how response strength increases after reward and decreases after nonreward. The one we have found that best combines simplicity and generality incorporates the organism's history of both reward and responding. It is not the last word.[4] Nevertheless, it shows how a simple moment-by-moment process can yield complex and often adaptive behavior. I show later that prospect theory, although not a theory in the same sense as the CE model—it is classification of the data, not a predictive model—nevertheless implies a similar underlying framework.

The assumptions of the *cumulative effects (CE) model* are as follows:

1. In a given choice situation, there is a repertoire of *candidates* (responses).
2. Each response has a certain *strength*, which is given by the following, equation 1:

$$X_i = \frac{\sum R_i + R_i(0)}{\sum x_i + x_i(0)}$$

where X_i is the strength of response i, R_i, and x_i are the rewards received and i-responses made; $R_i(0)$ and $x_i(0)$ are initial conditions. In words, the strength of each response tendency is just the cumulated payoff probability for that response, biased by prior reward and response totals. The learning rule is very simple: a response not followed by reward increases the denominator and thus reduces X_i; a reward increments the numerator proportionately more than the denominator and thus increases X_i.

3. Response tendencies compete to become active in winner-take-all fashion.

This simple model behaves rationally in choosing between two options that pay off with different probabilities: it will eventually fixate on the higher-probability choice.

More surprising is that it can duplicate pigeons' performance on successive-reversal learning (Davis et al. 1993; Staddon 2014b). In these

4. For one thing, it is deterministic, not stochastic. The randomness in my example is provided by the schedule. On fixed-ratio schedules, the result is usually indifference, which is probably not correct—although the necessary experiment does not seem to have been done.

experiments the pigeon chooses between two keys, as in the previous example. On day 1, pecks on the left key (say) are paid off randomly with probability 1/8; the right key is not paid off. On day 2, this is reversed: right p(reward) = 1/8, left p(reward) = 0.

And so on for many days: left and right rewarded on alternate days.

Unsurprisingly perhaps, pigeons improve over days, switching their preference faster and faster each day. This is called "reversal learning set," and many cognitive explanations have been offered for it. But the CE model can duplicate the effect because, over successive reversal days, the X values of the two choices get closer and closer so that switching from one to the other (because of the winner-take-all rule) gets faster and faster. No need to postulate a "set" or anything beyond simple learning.[5] The same model also predicts, correctly, that reversal performance will improve less if reversals take place every four days rather than every day.

So what does the CE model (and many models like it) predict should happen when the animal must choose between two *identical* ratio schedules? Maximization makes no prediction because the animal's choice has no effect on payoff. A version of H. A. Simon's satisficing[6] might predict random fixation on one side, given that switching entails some small cost, but that is not what happens.

Figure 1 shows a simulation of the CE model in this situation and some data (Horner and Staddon 1987).[7] The model predicts, and the data show, that the animal's behavior depends on the absolute value of the payoff probability. If it is high, the animal tends to fixate (randomly) on one or other choice; if it is low, it tends toward indifference.

Gordon Tullock (1971), who was one of the first economists to look at biological problems from an economic viewpoint, proposed something like this in his nifty little note "The Coal Tit as a Careful Shopper": "Assume that the bird, like any other food-consuming animal, explores a number of different food sources [pine cones with grubs, in this case], and then settles on the one which provides the most food with the least energy expended. Since it would tend to exhaust the food supplies in that particular area, it should be 'programmed' in such a way that it periodically investigates different areas." The CE model adds dynamics plus the prediction that the bird turns sooner to "periodical investigation" when it pecks a

5. Pigeons never show *spontaneous reversal*—switching on the first response each day. They always require a few unsuccessful responses to yesterday's "hot" key before switching.

6. See en.wikipedia.org/wiki/Satisficing.

7. For an extensive theoretical analysis of this and related situations, see Staddon 1988.

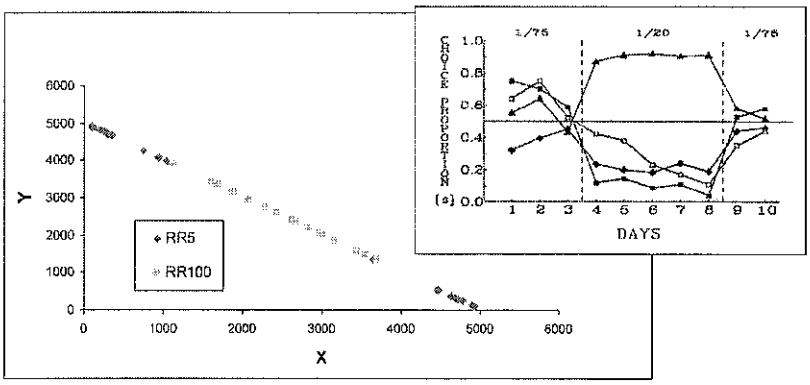


Figure 1. The pattern of choice between identical schedules as a function of absolute reinforcement rate. *Note:* CE model simulation: Two identical random ratios, 5 and 100. Preferences after twenty runs each of 4,993 time steps. Graph plots total responses: X vs. Y for two ratios: 5 and 100. Exclusive choice is favored at the smaller ratio (diamonds), indifference at the large value (squares). Initial values in Equation 1: $R_L = R_R = 5$, $x = y = 10$. Data: The figure plots the proportion of choices of the right-hand alternative across daily sessions for each of four pigeons for two equal-random-ratio conditions, $p = 1/75$ and $p = 1/20$, in ABA sequence (Horner and Staddon 1987, fig. 1). In both simulation and data, exclusive choice is favored at the high probability, indifference at the low.

relatively grub-free pine cone. This in outline is how one theoretical school of psychology thinks about the processes that underlie choice. How does prospect theory treat problems like this?

Human Choice

Animals must actually experience the different payoffs to arrive at a stable preference. Not so for people: you can just ask them (Kahneman 2011).[8] Consequently, the study of human choice behavior has developed along very different lines from the study of choice in animals. Nevertheless, I show that both can be conceived in a similar way.

8. But of course what they do when they actually experience outcomes may be very different from what they say in advance they will do. See, e.g., Ludvig, Madan, and Spetch 2014. Portions of this section, by the way, appear in Staddon 2016.

The study of human choice differs from animal choice in four main ways:

1. People begin with a stock of *wealth*. There is no real equivalent in animal experiments.[9]
2. There is no animal equivalent to *loss*.
3. Individual differences: *There is rarely unanimity in choice experiments with people*. Even when there is a statistically significant preference, 30–40 percent of subjects may deviate from the majority choice. In most animal choice experiments the effects are reversible (see figure 1), so the result can be replicated with the same subjects and no statistics are required. In practice, close to 100 percent of subjects give the same result.
4. Human choice patterns can be changed by experience: risk aversion can be changed to risk seeking by appropriate training or prompting,[10] for example. In most animal studies, choice is stable.

Because people will give immediate answers to many hypothetical choice questions, and because their decisions can often be accurately anticipated, hypotheses can be tested quickly. This strategy was adopted by two of the leaders in this field, Daniel Kahneman and the late Amos Tversky. They posed questions to themselves, and based on their own answers (which were later checked with groups of human subjects), they quickly found flaws in *utility theory*, the standard explanation in economics.

Prospect Theory

Daniel Bernoulli in the eighteenth century pointed out that the utility of a given increment in wealth is inversely related to the amount of wealth you already have. A hundred dollars is worth much more to a pauper (net wealth: $5) than to a millionaire.[11] Anticipating a now well-known principle in psychophysics, the Weber-Fechner law, Bernoulli proposed that the relationship between wealth and utility is logarithmic, so that equal

9. One experiment has tried, with some success, to induce monkeys to treat a stock of tokens as wealth: see Lakshminarayanan, Chen, and Santos 2011. This effect fits in with the more general analysis I present here, but it does suggest that monkeys are smarter than pigeons.

10. Experienced traders show fewer biases than novices; for example, see Kahneman 2011.

11. Comparing utility between individuals is epistemologically flawed, of course, even though it is the basis for the idea of social justice. We can verify that person A prefers X to Y and person B the reverse. We cannot say whether A values X more than B does. After all, the millionaire may be one just because he values a dollar, any dollar, more than does the pauper. Bernoulli's assumption is verifiable for individual utility but not for interpersonal comparison.

ratios correspond to equal value. A pay raise of $1,000 is worth as much to an employee whose base pay is $10,000 as a raise of $10,000 to one who makes $100,000.

The standard utility function is negatively accelerated, which means diminishing marginal utility. Because each increment of value adds a smaller increment of utility, people will generally be *risk averse* with respect to gains because doubling a reward less than doubles its utility. For example, in a typical experiment (Problem 3 in Kahneman and Tversky 1979),[12] ninety-five people were asked to choose between two outcomes: 4,000 with probability 0.8, versus 3,000 for sure. Eighty percent chose the sure thing, even though it is less than the expected value of the gamble: 3,000 < 3,200. Apparently, 0.8 times the utility of 4,000 is less than the utility of 3,000, $0.8 \times U(4{,}000) < U(3{,}000)$, because the utility function is negatively accelerated.

Yet expected value alone explains Problem 4, where 65 percent of the same ninety-five people preferred a 0.2 chance of 4,000 to a 0.25 chance of 3,000: in expected values, 800 was now preferred to 750—no sign of diminishing returns there. Evidently *certainty* has an effect that goes beyond the probability of one.

There is another way to look at Problem 3. The "3,000 for sure" option is guaranteed. It increments your current state of wealth, *unless* you choose the risky option, in which case, with probability 0.2 (the chance you *do not* win the 4,000), your wealth could go down by 3,000, a *loss*. So the issue becomes the following: is $|0.2 \times U(-3{,}000)| > U(4{,}000 - 3{,}000)$, that is, is 0.2 times the cost of losing 3,000 (i.e., actual expected value 600), greater than the benefit of a net gain of 1,000? And the answer is yes if the origin of the utility curve is shifted up by 3,000 on the utility axis (i.e., 3,000 is subtracted from each utility). In this case the disutility of a loss of 600 is greater than the utility of a gain of 1,000. But such a shift requires a further assumption about how people *perceive* a choice situation. The shift of origin is not really required for Problem 3, since the all-positive utility curve works as well, but it is necessary to account for other results.

Data like these led Kahneman and Tversky to propose *prospect theory* as a replacement for standard utility theory. Prospect theory is a

12. The example is Problem 3; amounts are in Israeli currency. This version of prospect theory was elaborated in Tversky and Kahneman 1992, and there is a vast secondary literature. See also Kahneman's readable book *Thinking Fast and Slow* (2011) and his Nobel Prize address: www.nobelprize.org/mediaplayer/?id=531.

hybrid. The "front end" is what they term *editing*: "Prospect theory distinguishes two phases in the choice process: an early phase of editing and a subsequent phase of evaluation" (Kahneman and Tversky 1979, 274). Editing is their name for cognitive processes that set up the problem for the *evaluation* part of the theory. Editing is a process, but it is not defined in any calculable way. Terms such as framing, combination, segregation, isolation, coalescing, and cancellation have been used to label operations that the decision maker can use to simplify and structure the choice problem so that it can be evaluated. Exactly when these processes are invoked and precisely how they work are not specified by any economic theory.

The core of prospect theory is the second phase: *evaluation*, which is a modified version of expected utility theory. Evaluation is a teleological/functional theory; it describes an outcome but not the process the system uses to achieve it. Evaluation, according to prospect theory, modifies utility theory in three ways:

1. The idea of a *reference point* or adaptation level, which is set by the subject's current state of wealth.[13] The reference point moves the origin of the utility curve to a point representing the subject's current wealth. A choice option that yields a state of wealth less than the reference point is perceived as a loss. The idea that people are sensitive to changes rather than absolute values is a core principle of perceptual psychology.
2. Thus, the upper-right quadrant of the utility graph is the same as before. But the lower left part is modified, to deal with the following results involving losses. There are three steps; the last two require changes in the standard form. First, most people will choose a 50 percent chance of losing $100 only if the other possibility is a win of $200 or more: "People tend to be risk averse in the domain of gains and risk seeking in the domain of losses" (Kahneman 2011, 344). Thus, the first part of the southwest quadrant must be steeper than the first part of the northeast. This requires no big change in the shape of the standard Bernoulli utility graph, which gets steeper as it approaches the origin.

But, second, 65 percent of subjects preferred a 0.2 chance of winning 4,000 over a 0.25 chance to win 3,000 (expected values: 800 > 750;

13. See also Markowitz 1952.

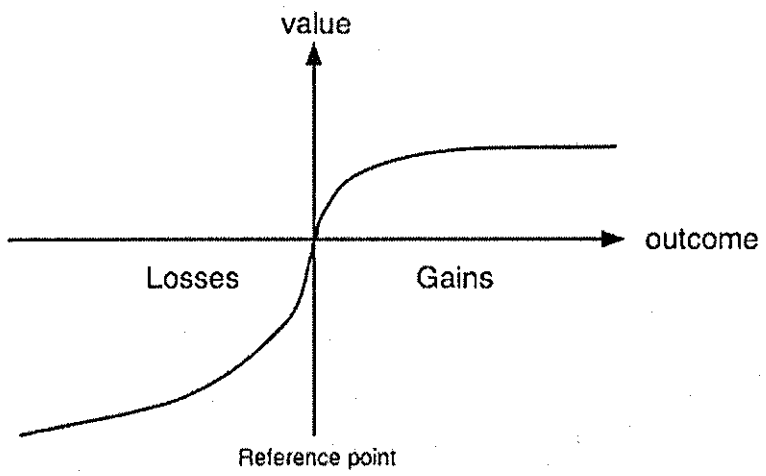


Figure 2. Prospect theory utility function

Problem 4), that is, they were in line with expected value theory and not risk averse. So the curve must straighten out at intermediate probabilities. And finally, 86 percent of people preferred a 0.9 chance to win 3,000 over a 0.45 chance to win 6,000 (Problem 7), even though both gambles have the same expected value. Now subjects showed risk aversion. Hence the curve in the southwest quadrant must begin steep, then straighten, and then flatten out. This is the iconic prospect theory utility graph (see figure 2: Kahneman and Tversky 1979, fig. 3).

3. Gambles at extreme odds do not fit even this modified utility graph. For example, 73 percent of people prefer a .001 chance to win 6,000 over a .002 chance to win 3,000 (Problem 8), even though the expected values are the same, that is, back to risk seeking.

Another exception is provided by something called probabilistic insurance. Suppose that you are indifferent about whether or not to insure your house against earthquakes. A creative salesman then makes you this offer: we are willing to insure you for just 45 percent of the original premium if we are allowed to toss a coin after an incident to decide whether we pay you or just refund your premium. Obviously this is a better deal than the original one, which you were on the point of accepting. Yet people will

usually reject it. Not that they are against probabilistic insurance on principle. As Kahneman and Tversky point out, the decision to quit a risky habit like smoking is a form of probabilistic insurance.

These problems cannot be solved through changes in the utility curve alone. To fix them, Kahneman and Tversky (1979, fig. 4) introduce a somewhat untidy transformation of probabilities ("π is not well-behaved near the end-points"), termed *decision weights*, that parallels Bernoulli's transformation of value. Evidently both the utility curve and the probabilities must be transformed to accommodate data.

A Common Framework

Attempts to quantify all these modifications require four or even five fitted parameters.[14] That is just too many for a useful scientific theory. John von Neumann, genius, computer pioneer, and coinventor of game theory, famously remarked: "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk." Prospect theory has evolved into a set of descriptions, not a predictive system. Nevertheless, as I will show, the underlying concepts can be related in an interesting way to choice in nonhuman animals.

Human choice behavior is affected by many things: not just the numbers you are presented with but your state of wealth; your experience, both recent and historical; the kind of goods on offer—money or things, for example—the context in which the task is presented; and your willingness to think through the options.

Kahneman and Tversky tackle this complexity by pointing out that prospect theory deals only with decisions that are made quickly. They distinguish between fast and slow cognitive processing,[15] what Kahneman calls System 1 (fast, thoughtless) and System 2 (slow, reflective). Presumably, the rapid process is simpler to understand than the slow one.

Their scheme is not as different from the analysis of animal choice behavior that I presented earlier as you might think. Consider three aspects of prospect theory: the role of consciousness, framing, and the fast-slow distinction.

14. See, e.g., Wilkinson and Klaes 2012, 163.

15. I think that the fast-slow distinction is actually due to Kahneman, but because Tversky would probably have agreed had he lived, and because prospect theory is identified with both of them, I use both names.

Prospect theory is usually considered a cognitive, as opposed to a behavioristic, system. Yet the role of consciousness, especially in the fast system, is minimal: "The mental work that produces impressions, intuitions, and many decisions goes on in silence in our mind"; and "Studies of priming effects have yielded discoveries that threaten our self-image as conscious and autonomous authors of our judgments and our choices" (Kahneman 2011, 4, 55). Compare this with the following comment by a behaviorist: "Astrophysicists now tell us that more than 80 percent of the matter in the universe, so-called *dark matter*, cannot be observed. The unconscious is the dark matter of psychology. Its processes are responsible for all creative activity and most recollection" (Staddon 2014a, 9). The autonomy of the unconscious is not even a very new idea. Well before Sigmund Freud, the eccentric genius Samuel Butler (1835–1902) referred frequently to the unconscious in his posthumously published novel *The Way of All Flesh*, which is in effect an autobiographical study of his own psychological development.

Cognitive psychologists are admittedly more interested than behaviorists in consciousness. But the contemporary view, like Butler's, is that consciousness is not an active agent but something like a workspace, which "allows the novel combination of material" (Baddeley 2007). Let us agree, therefore, that cognitivists and behaviorists no longer differ greatly on the role of the unconscious.

Framing is the term Kahneman and Tversky give to the effect of context and the way a problem is presented on the response that is made. The term does not occur in the 1979 paper but, with a number of other labels such as nonlinear preference and source dependence, was made necessary by new data that did not fit the original formulation. An example of framing is "The statement that 'the odds of survival one month after surgery are 90%' is more reassuring than the equivalent statement that 'mortality within one month of surgery is 10%.' Similarly, cold cuts described as '90% fat-free' are more attractive than when they are described as '10% fat.' The equivalence of the alternative formulations is transparent, but an individual normally sees only one formulation, and what she sees is all there is" (Kahneman 2011, 88). The way in which a question is asked (this is also sometimes termed *choice architecture*) can have a huge effect on the subject's response.

There is a counterpart to framing in animal choice. It grows out of the consensus that learning in animals is best conceived as a process

of variation and selection, that is, selection by consequences, and variation constrained by the situation:

> For example, Pavlovian conditioning, which allows a neutral stimulus to acquire signal properties, will itself give rise to a repertoire of reinforcer-related activities from which operant reinforcement can then select. A stimulus associated with food, or food by itself, will induce a wide range of food-related activities in a hungry animal—activities from which operant contingencies can select. Pigeons peck, chickens peck and scratch, raccoons manipulate. Pavlovian conditioning, with (say) a food US [unconditioned stimulus], in effect frames or *labels* the context as food-related. The label then limits the emitted behavior to a food-related repertoire, which is defined partly by past history but also by the organism's evolutionary history. (Staddon 2014a, 92)

The situation-induced behavioral repertoire of animals is constrained—framed—in much the same way as human choice behavior. Sometimes "Pavlovian framing" will facilitate learning—if the desired response is within the induced repertoire; but sometimes it will not—as when the induced behavior interferes with the trainer's target behavior (so-called instinctive drift).

Note that framing in this sense means that all rewards are not equivalent—as most economists seem to assume. The repertoire induced by money will be different from the repertoire induced by love of ideas or a wish to cure patients. Consequently, paying teachers or doctors more will not necessarily produce better teaching or medical care.

Finally, let us look at Kahneman's fast-slow distinction. He contrasts the quick answers he and Tversky got to their choice questions—answers that were often "irrational" in the sense that they did not maximize gain—with the slower and more "correct" answer that most people arrive at after deliberation. The quick system responds with answers that are more "available" or "accessible" than other responses that may in fact be better. Kahneman (2011, 415) also tells his readers to "remember that the two systems do not really exist in the brain or anywhere else." They are ways to talk about the fact that people may respond one way at first and another way after given some time for reflection.

But there is no difference between Kahneman's System 1 and System 2 and what I earlier called "active" and silent" responses. The active response is the first thing to occur in a given situation. But if the conse-

quences are unsatisfactory, the active response is eventually supplanted by another response from the "silent" repertoire—Kahneman's "slow" system. Individual differences reflect available response repertoires of different individuals—differences that reflect different histories and propensities. Human responses are more complex than key pecks, of course. They are heuristics, rules of thumb and analytic strategies, the "adaptive toolbox." But there is no reason to doubt that they are emitted and selected by consequences, just like the simpler response of pigeons.

The economic approach has been looking for utility functions and decision weights that best fit a set of quite variable choice data. Since these data usually depend on historical factors that are not part of the analysis, this is likely to be an endless quest. A symptom of the difficulties this causes is that many very clever people have puzzled over the fact that human behavior often violates logical axioms such as transitivity and independence.

Transitivity is simple: if A is preferred to B and B to C, then logic says that A must be preferred to C—but often people disagree.

Independence is a bit more complicated. Look at table 1, which shows two pairs of gambles, A versus B and C versus D (Kahneman and Tversky 1979, Problem 1). Boldface shows which one of each pair is generally preferred (again, there are always individual differences): B over A, but C over D. There are two things to notice about this pattern: first, people pick the (slightly) lower expected value in the first pair but the (slightly) higher EV in the second. Second, the only difference between these two pairs of gambles is that the payoff .66*2400 (italics) has been added to both choices in A and B. Otherwise, the choice between A and B is the same as between C and D. Thus, most people's choices violate the independence assumption. The prospect theory explanation is the primacy of the certain outcome (Gamble B), which describes the result but does not explain it.

But why on earth should violations of logical axioms like this be worrisome and consume so much mental energy? Behavior is not physics or predicate calculus. Even human beings rarely maximize in an explicit way, following all the rules of optimality theory—especially when they must decide quickly. They do not consider all the options in an unbiased, neutral, and history-independent way, and then compute the costs and benefits and choose accordingly. There is therefore *no reason whatever* that human (and animal) choice should fit any set of simple axioms.

Table 1 Violation of the Independence Assumption **Preferences** Change When a Constant Is Added to Each Choice

	Probability			Expected Value
Gamble	0.33	0.01	0.66	
A	2500	0	2400	2409
B	2400	2400	2400	2400
C	2500	0	0	825
D	2400	2400	0	816

Conclusion

The problem these examples illustrate goes beyond prospect theory. It is a fundamental problem for any outcome-based theory like expected utility. EUT is a functional/optimality theory and not a mechanism. It says that behavior will adjust so that marginal utilities are matched (or whatever), but provides no causal account for how this is accomplished. But there must be a causal account, a mechanism, and that mechanism will fail to optimize—to "act rationally"—under some conditions. Hence EUT will also fail sometimes—as Kahneman and Tversky eventually recognized. They did not, I think, understand the causal-functional distinction. They thought prospect theory was what I call a real (i.e., causal) theory and so were surprised when they had to keep adding kludges to make it work with a wider range of data.

Functional—mechanism-free—theories are not without value, but they will always run up against limits. Theoretical behaviorism, based on the simpler problems posed by learning in animals, suggests that it may be more profitable to look at the historical factors that determine people's response repertoire. Rather than rely on some form of utility theory as the ultimate explanation for economic behavior, future research might better examine the historical and cultural—*causal*—factors that explain why a particular question format drives people to frame a problem in a certain way, a way that allows them to pick "best in repertoire" even if it is not optimal.

Exactly what are the "editing" processes that cause people to perceive logically identical problems quite differently? What kind of training affects how decision makers perceive different kinds of problems? The optimality approach—marginal utility theory—can help us understand a particular choice problem. It will never tell us how people actually choose.

References

Baddeley, A. 2007. *Working Memory, Thought, and Action.* www.oxfordscholarship.com/view/10.1093/acprof:oso/9780198528012.001.0001/acprof-9780198528012-chapter-1.

Davis, D. G. S., J. E. R. Staddon, A. Machado, and R. G. Palmer. 1993. "The Process of Recurrent Choice." *Psychological Review* 100: 320–41.

Gigerenzer, G., and R. Selten, eds., 2001. *Bounded Rationality: The Adaptive Toolbox.* Cambridge, Mass.: MIT Press.

Horner, J. M., and J. E. R. Staddon. 1987. "Probabilistic Choice: A Simple Invariance." *Behavioural Processes* 15: 59–92.

Kahneman, D. 2011. *Thinking Fast and Slow.* New York: Farrar, Straus and Giroux.

Kahneman, D., and A. Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47 (2): 263–91.

Lakshminarayanan, Venkat R., M. Keith Chen, and Laurie R. Santos. 2011. "The Evolution of Decision-Making under Risk: Framing Effects in Monkey Risk Preferences." *Journal of Experimental Social Psychology* 47: 689–93.

Ludvig, E. A., C. R. Madan, and M. L. Spetch. 2014. "Priming Memories of Past Wins Induces Risk Seeking." Online first publication, December 22. doi:10.1037/xge0000046.

Markowitz, H. 1952. "The Utility of Wealth." *Journal of Political Economy* 60: 151–58.

Staddon, J. E. R., ed. 1980. *Limits to Action: The Allocation of Individual Behavior.* New York: Academic Press.

———. 1988. "Quasi-dynamic Choice Models: Melioration and Ratio-Invariance. *Journal of the Experimental Analysis of Behavior* 49: 303–20.

———. 2007. "Is Animal Learning Optimal?" In *Constructal Theory of Social Dynamics*, edited by A. Bejan and G. W. Merkx, 161–67. New York: Springer Verlag.

———. 2012. *The Malign Hand of the Markets: The Insidious Forces on Wall Street That Are Destroying Financial Markets—and What We Can Do about It.* New York: McGraw-Hill.

———. 2014a. *The New Behaviorism.* 2nd ed. Philadelphia: Psychology Press.

———. 2014b. "On Choice and the Law of Effect." *International Journal of Comparative Psychology* 27 (4): 569–84.

———. 2016. *Adaptive Behavior and Learning.* 2nd ed. New York: Cambridge University Press.

Tullock, Gordon. 1971. "The Coal Tit as a Careful Shopper." *American Naturalist* 105: 77–80.

Tversky, A., and D. Kahneman. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty* 5: 297–323.

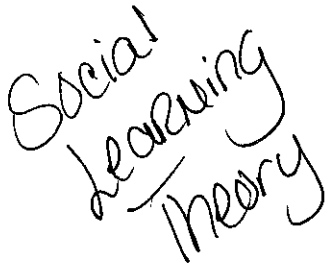
Wilkinson, N., and M. A. Klaes. 2012. *Introduction to Behavioural Economics.* 2nd ed. Basingstoke, U.K.: Palgrave Macmillan.

J Autism Dev Disord (2017) 47:615–625
DOI 10.1007/s10803-016-2983-2

ORIGINAL PAPER

Trust and Deception in Children with Autism Spectrum Disorders: A Social Learning Perspective

Yiying Yang[1] · Yuan Tian[2] · Jing Fang[3] · Haoyang Lu[4,5] · Kunlin Wei[1] · Li Yi[1]



Published online: 16 December 2016

Abstract Previous research has demonstrated abnormal trust and deception behaviors in children with Autism Spectrum Disorders (ASD), and we aimed to examine whether these abnormalities were primarily due to their specific deficits in social learning. We tested 42 high-functioning children with ASD and 38 age- and ability-matched typically developing (TD) children in trust and deception tasks and a novel condition with reduced social components. Results indicated that while TD children improved their performance with more social components, children with ASD lacked this additional performance gain, though they performed similarly as TD children in the condition with reduced social components. Our findings highlight that deficits of ASD in trust and deception are primarily associated with failure of use of social cues.

Keywords Autism Spectrum Disorder · Trust · Deception · Social learning

Introduction

Humans can benefit from social information during social interaction, such as eye gazes and facial cues, to learn social standards (e.g., Bushwick 2001; Hareli et al. 2015), rules (e.g., Jones et al. 2013; Wang et al. 2015), and language (e.g., Baldwin 1993; Dominey and Dodane 2004; Tomasello and Farrar 1986; Yu and Ballard 2007). Social learning is a cumulative and cyclic process, which begins almost at birth and takes several years of training and practice (Bushwick 2001). However, individuals with autism spectrum disorder (ASD), who have limited attention to and understanding of social cues, may be less likely to benefit from social learning. This impairment has been proposed to contribute to core symptoms of ASD, such as defective social communication (Bushwick 2001). In the current study, we investigated two kinds of social behaviors, trust and deception, under the framework of social learning.

Trust and deception are two types of social behaviors that children learn as they grow up. Preschoolers develop selective trust to determine whether others are trustworthy based on their past reliability, motives, and past deceiving behaviors (e.g., Corriveau and Harris 2009; Koenig and Harris 2005; Vanderbilt et al. 2011). Children with ASD, however, tend to show a trust bias towards other people's testimony, and unconditionally give credence to information provided by them (Yi et al. 2013, 2014). They also experience difficulty using facial cues to selectively trust others (Ewing et al. 2015). Closely related to trust behaviors, deception, a milestone of children's social cognitive development, also experiences

✉ Kunlin Wei
wei_kunlin@126.com

✉ Li Yi
yilipku@pku.edu.cn

1 School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, 5 Yiheyuan Road, Beijing 100871, China

2 Southern China Research Center of Statistical Science, School of Mathematical and Computational Science, Sun Yat-sen University, Guangzhou 510275, China

3 Qingdao Autism Research Institute, 7 Jinsong Seven Road, Qingdao 266000, Shandong, China

4 Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China

5 Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China

remarkable development in the preschool years. As children grow up, they deceive others more frequently with more sophisticated strategies (Chandler et al. 1989; Russell et al. 1991; Sodian et al. 1991). However, children with ASD tend to have difficulties with lying, deceiving, and manipulating beliefs of other people (Baron-Cohen 1992; Li et al. 2011; Russell et al. 1991; Sodian and Frith 1992; Talwar et al. 2012; Yirmiya et al. 1996).

Typically, children's trust and deception performance was assessed in the context of social learning, where children need to rely on social cues provided by another person (Yi et al. 2014). Yi et al. (2014) found that children with ASD display slower learning rates than TD children regarding how to distrust and deceive an adult after being repeatedly deceived. Given the complexity of the tasks, it remains an open question whether these learning impairments primarily stem from deficits in general learning abilities as opposed to social learning per se. The present study aimed to examine whether similar learning impairments persist with reduced social components when cueing information is provided by physical markers as opposed to by a human actor. We compared the trust and deception behaviors between ASD and TD groups in a social cue condition, adopted from tasks in Yi et al. (2014), as well as in a non-social cue condition in which children had to rely on physical cues. In the social cue condition, children had to find a hidden prize, with an informant always providing misleading information about its location; then they switched the roles: the informant tried to find the prize and children were given a chance to help or misguide her. The non-social cue condition followed a similar procedure as the social cue condition except that the informant, an actual human in the social cue condition, was replaced with a physical marker in the non-social cue condition. Compared to the social cue condition, all the executive steps are still present in this non-social cue condition: children need to learn whether the cueing information is false, inhibit impulsiveness to follow the false information, and suppress the action to indicate the truth. However, they do not need to rely on social cues provided by another person (e.g., verbal instructions and gestural information), which requires understanding others' intention to distrust and manipulate others' mental states to deceive. Thus, if children with ASD have relatively intact general learning, they should have similar performance as TD children in the non-social cue condition. On the other hand, based on their specific deficit in social learning, we would expect a group difference in the social cue condition. On the contrary, if they were impaired in general learning, they should display poorer performance in comparison to TD children in both conditions.

Method

Participants

Forty-two Chinese children with ASD (age range 4.31–8.22 years, $M=5.95$, $SD=0.84$) and 38 age- and ability-matched Chinese TD children (age range 4.00–7.00 years, $M=5.73$, $SD=0.71$) were recruited for this study. This age range was chosen based on the previous study (Yi et al. 2014) and our pilot results, to ensure that most children, especially children with ASD, could understand the task instructions and complete the task. To determine the minimum sample size required in the current study, we have conducted a prior power analysis based on effect sizes discovered in a previous study with similar paradigms (Yi et al. 2014). Our sample size was adequate for obtaining a power over 0.9. None of the participants had participated in our previous studies or relative research projects.

Children in the ASD group were recruited from a special school for ASD, while children in the TD group were recruited from a normal primary school in China to participate in the study. All children with ASD included in our study were previously diagnosed by two experienced pediatric clinicians strictly according to the diagnostic criteria for ASD in the DSM-IV-TR (American Psychiatric Association 2000). Since the Autism Diagnostic Observation Schedule (ADOS; Lord et al. 2000) and the Autism Diagnostic Interview-Revised (ADI-R; Lord et al. 1994) have not been officially validated and widely used in China, we confirmed the diagnosis of children with ASD using the Chinese version of Autism Spectrum Quotient: Children's version (AQ-Child[1]), the Social Responsive Scale (SRS[2]), and the Social Communication Questionnaire (SCQ[3]). All questionnaires were scored based on the parents' reports.

Children with ASD were recruited first and randomly assigned to either the social cue or the non-social cue condition. TD children were then recruited to match the ASD group by their (a) chronological age, (b) non-verbal IQ, measured by Combined Raven Test (CRT-C2), and (c) verbal ability, measured by Peabody Vocabulary Test-Revised (PPVT-R) (see Table 1). TD children were then also randomly assigned to the social cue or the non-social cue condition.

[1] The AQ was translated by R.W.S. Chan, W.S. Liu, K.K. Chung, C.S. Sheh & E.K.F. Woo (2008), Hong Kong Working Group on ASD, cprc@hkusua.hku.hk.

[2] The SRS is the Chinese version translated from the Constantino and Gruber (2002) version.

[3] The SCQ is the Chinese version translated from Rutter et al. (2003) Copyright© 2003 by Western Psychological Services (Fourth Printing, April 2012).

Table 1 Sample sizes, age, CRT scores, and PPVT scores of the four groups of participants

Group	Social cue condition		Non-social cue condition	
	TD	ASD	TD	ASD
N (female)	17 (5)	22 (4)	21 (6)	20 (1)
Age range	4.00–7.00	4.62–7.62	4.15–6.96	4.31–8.22
Mean age	5.72 (0.69)	6.09 (0.76)	5.73 (0.75)	5.78 (0.91)
CRT (raw scores)	25.41 (8.85)	27.27 (9.01)	25.24 (5.77)	25.95 (9.65)
CRT (standardized scores)	100.69 (13.51)	100.59 (16.25)	100.00 (9.25)	100.69 (18.22)
PPVT	81.29 (11.21)	74.14 (12.35)	83.33 (14.50)	74.55 (15.29)
VMA[a]	6.29 (0.59)	6.05 (0.65)	6.48 (0.60)	6.00 (0.80)

[a] *VMA* verbal mental age, standardized scores calculated from PPVT scores

Procedure

Distrust and Deception Tasks

A 2 (participant group) * 2 (condition) between-subject design was adopted for the experiment. We asked children to randomly participate in either the social cue or the non-social cue condition, instead of completing both conditions, considering the potential carry-over effect, practice effect, and order effect. Each child completed the distrust and the deception tasks sequentially.

The child, Experimenter 1 (E1), and Experimenter 2 (E2) each sat at one side of a table where three identical boxes were placed along a line. The child was asked to participate in a hide-and-seek game to win prizes with the experimenters, and was encouraged to find "as many prizes as possible." The experimenters were trained sufficiently beforehand on performing the tasks: they were asked not to show any expressions while they interacted with children. We used the same experimenters throughout the whole experiment to minimize the potential impact of different experimenters on children's behavior. The whole procedure was illustrated by the schematic drawing in Fig. 1.

Social Cue Condition

In the distrust task, E1 showed three identical boxes and explained the rules to the child: (a) E1 hid a prize in one of the three boxes and the child had to guess where it was; (b) E2 then examined all the boxes and showed the child where the prize was by pointing at one box; (c) the child could keep the prize if he successfully found it, otherwise E2 would keep it.

After confirming that the child understood the rules, E1 announced that the game began: (a) both the child and E2 turned their backs to the table, then E1 kept one box empty, and placed a prize in each of the other two boxes; (b) the child and E2 faced the table again, then E2 examined all the boxes, pointed at the only empty box, and said to the child "the prize is here"; (c) E1 asked the child to pick one box for the prize; (d) E1 opened the box chosen by the child and provided the feedback (i.e., whether or not the child received the prize for his behavior on each trial) accordingly; (e) if the child did not choose the empty box, E1 gave him the prize; otherwise, E1 took a prize out of either one of the two remaining boxes and gave it to E2. The task was repeated for ten times. Finding the prize in five consecutive trials was considered successfully passing the task. Then the session stopped, and the remaining trials were marked as correct.

Following the distrust task, the child proceeded with the deception task, in which the child and E2 essentially exchanged their roles. First, E1 explained the new rules of the deception task to the child: (a) the child should hide a prize in one box for E2 to look for; (b) E2 could keep the prize if she was correct; otherwise, the child could keep it; (c) before taking a guess, E2 asked the child about the location of the prize, and the child should respond by pointing at one box. After confirming that the child understood the rules, E1 announced that the game began: (a) E2 turned her back to the table, and the child hid a prize in one box; (b) E2 turned around to face the table and asked the child about the whereabouts of the prize, and the child responded; (c) E2 "guessed" that the prize was hidden in the box that was just pointed to by the child; (d) E1 opened the pointed box and provided feedback (i.e., whether or not the child received the prize for his behavior on each trial) accordingly; (e) the child received the prize if he pointed at an empty box; if the child pointed at the box with prize, E2 received the prize. Similar to the distrust task, the procedure was repeated for ten times. Children were considered as passing the task if they obtained the prize in five consecutive trials.

Non-social Cue Condition

In the non-social cue condition, we attempted to reduce social components of the distrust and deception tasks. Children no longer played a zero-sum game against E2. Instead, in the distrust task, they tried to win the prize by guessing

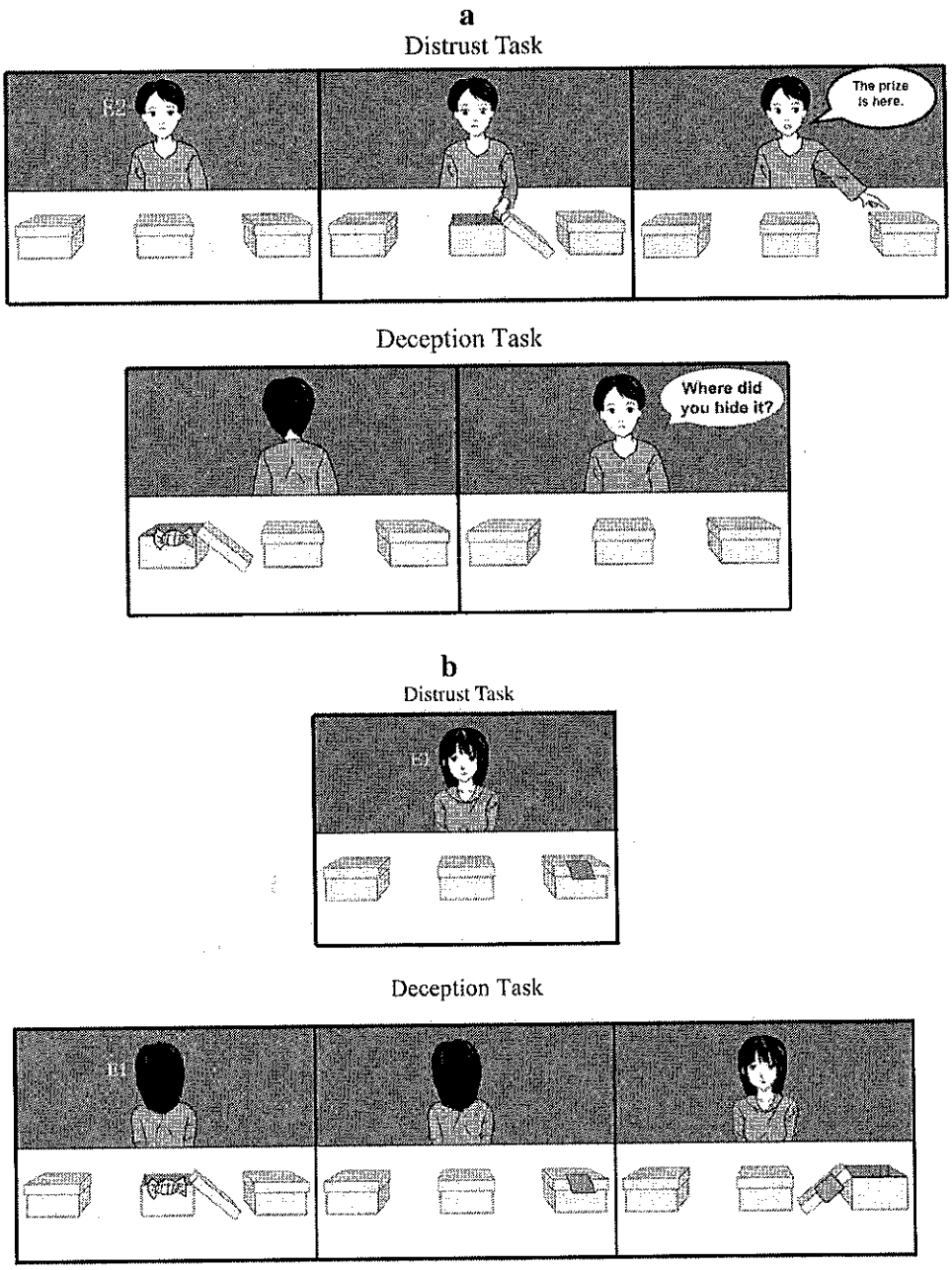


Fig. 1 Schematic drawing of the experimental procedures of the distrust and the deception tasks in the social cue (**a**) and the non-social cue (**b**) conditions

the location of the prize according to a removable sticker attached to one of the three boxes; in the deception tasks, they tried to win the prize by attaching a sticker to an empty box. Since E2 did not participate in the non-social cue condition, the child and E1 sat opposite to each other at a table.

In the distrust task, E1 showed three identical boxes to the child and explained the rules to him: there was a prize in one of the three boxes and the child had to guess where the prize was; he could keep the prize if he guessed right.

The remaining procedures were similar to those of the social cue condition, except that: (a) when the child turned his back to the table, E1 kept one box empty and placed a sticker on it, and placed a prize in each of the other two boxes; (b) right after the child faced the table again, E1 asked him to pick one box for the prize; (c) if the child chose the unbaited box, E1 gave him the prize; otherwise, if the child chose the baited box, E1 took a prize out of either one of the remaining boxes, put it aside, and stated that the prize belonged to neither the child nor E1.

Following the distrust task, the child proceeded with the deception task. E1 explained the new rules to the child: (a) the child should hide a prize in one box and then attach a sticker to one box; (b) the child could keep

Table 2 The average score over ten trials of the social cue and non-social cue conditions in the distrust and deception tasks

Task	Condition	TD	ASD
Distrust	Social cue	9.12 (0.70)	6.32 (2.32)
	Non-social cue	7.90 (1.37)	7.75 (1.65)
Deception	Social cue	9.71 (0.47)	6.95 (3.54)
	Non-social cue	9.00 (1.45)	6.85 (3.36)

the prize if he placed the sticker correctly; otherwise, neither the child nor E1 could keep the prize.

The remaining procedures were similar to those of the social cue condition, except for the following steps: (a) when E1 turned her back to the table, the child hid a prize and placed a sticker to one of the boxes; (b) E1 never asked the child about the whereabouts of the prize; (c) if the child placed the sticker on a box without the prize, he could keep the prize; otherwise, E1 put the prize aside and stated that it belonged to neither the child nor E1.

For the distrust task, if the child chose the box pointed by E2 or baited with a sticker, the trial was scored as 0; otherwise, it was scored as 1. For the deception task, if the child pointed or put a sticker on the box in which he hid the prize, the trial was scored as 0; otherwise, it was scored as 1. Hence, task performance was indicated by an array of 0 and 1 scores.

Data Analysis

We first averaged the task performance across ten trials and conducted ANCOVAs to compare these average scores between groups and conditions, with IQ and PPVT scores as the covariates. We used ANCOVAs since children's IQ and PPVT were correlated with overall trust and deception performance.[4] In our design, children "learned" from the feedbacks whether they received the prize on a trial-by-trial basis. To characterize their learning process, we also compared the two groups with two types of trial-by-trial analyses. One is to compare the two groups regarding the percentage of children who won the prize at each trial. The statistical test appropriate for this type of data is Fisher's Exact test. The other is the survival analysis that quantifies the percentage of children who eventually passed the task (i.e., winning the prize in five successive trials) and the speed of the learning process (i.e., how many trials needed to pass the task). Log-rank tests were used for group comparisons of the learning speed. Finally, we tested the correlations between the trust and deception performance (averaged scores) and their social ability in children with ASD.

[4] Specifically, the overall distrust performance was correlated with the IQ and PPVT scores, p's < 0.05; the overall deception performance was correlated with the PPVT score, $p < 0.01$.

Results

Distrust Task

Overall Performance

An ANCOVA was performed on the average scores of the distrust task (see Table 2) with experimental condition and participant group being independent variables, and IQ and PPVT scores being covariates. Results showed a significant main effect of the participant group: children with ASD were more likely to choose the baited box than TD children in the distrust tasks, $F(1,74) = 14.58$, $p < 0.001$. The main effect of the experimental condition was not significant, $F(1,74) = 0.14$, $p = 0.71$. However, the interaction between the participant group and the experimental condition was significant, $F(1,74) = 14.73$, $p < 0.001$. The simple main effect analysis indicated that in the social cue condition children with ASD were more likely to choose the baited box than TD children, $F(1,74) = 28.51$, $p < 0.001$, but there was no significant difference between the two groups in the non-social cue condition, $F(1,74) = 0.02$, $p = 0.88$.

Trial-by-trial Analysis

Trial-by-trial analyses were performed based on each group's performance in each trial (Fig. 2). The binary distribution (score 1 or 0) for each trial was compared between groups. In the social cue condition, participants in both groups performed similarly in the first trial, $p = 0.22$ (Fisher's Exact test, two-tailed), but TD children outperformed children with ASD on the second trial, $p = 0.01$ (Fisher's Exact test, two-tailed), and in most of the following trials. In the non-social cue condition, the two groups of children had similar performance in all trials, p's > 0.05. So, such a group difference in learning to distrust disappeared in the non-social cue condition when the social components of the tasks were reduced.

Survival Analysis

We conducted a survival analysis to examine how children's trust performance would be affected by feedback over the ten trials. The curves in Fig. 3 represent the percentages of children who failed to distrust E2 at specific time points, and the shaded areas represent 95% confidence intervals. Results of the log-rank tests revealed a group difference of the learning curves in the social cue condition:

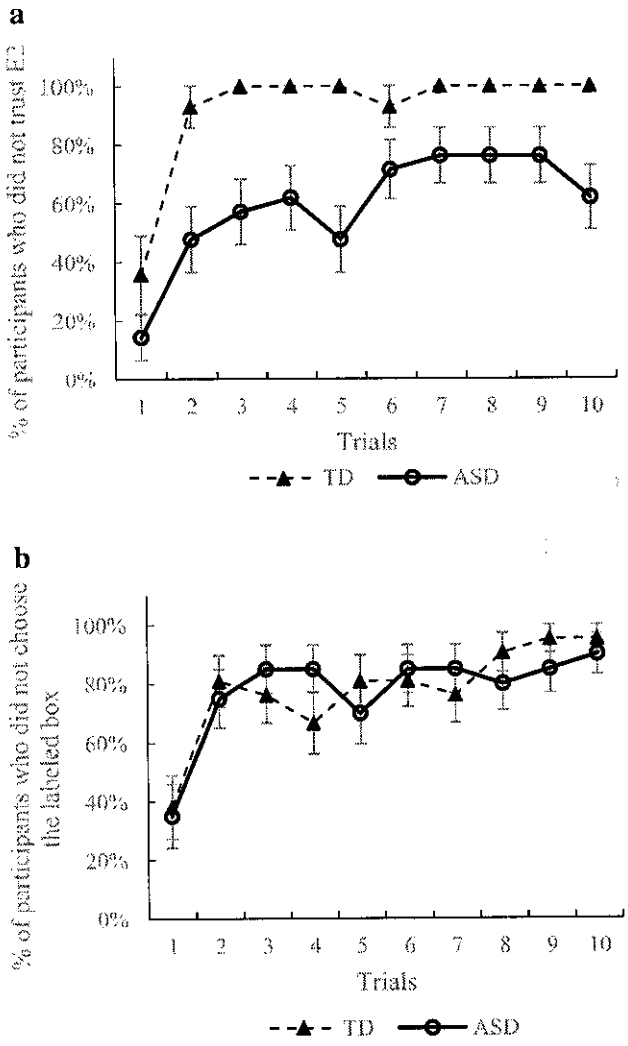


Fig. 2 Percentage of children with ASD and TD children who successfully found the prize as a function of trial numbers in the social cue (**a**) and the non-social cue (**b**) conditions of the distrust tasks. *Error bars* represent 95% confidence intervals

TD children learned faster than children with ASD in the social cue condition, $\chi^2(1)=14.3$, $p<0.001$. More specifically, the learning curve of the TD group showed a steep downward trend: on average 85.71% TD children began to choose the unbaited box in the first two trials, and 92.86% TD children successfully learned to choose the unbaited box at the end of the game. The ASD group had a less steep learning curve than the TD group, and only 47.62% children with ASD successfully learned to choose the unbaited box by the end of the game. In contrast, this group difference disappeared in the non-social cue condition: the two groups had similar learning rates, $\chi^2(1)=0.2$, $p=0.687$. At the end of the game, 51.74% TD children and 65% children with ASD successfully learned to find the prize. Assessing the difference between the conditions for each group separately, we found that TD children learned faster in the social cue condition than in the non-social cue condition, $\chi^2(1)=9$, $p=0.003$, while children with ASD had similar learning rates in both conditions, $\chi^2(1)=1.5$, $p=0.213$.

Deception task

Overall Performance

An ANCOVA was performed on the average scores of ten trials (see Table 2) of the deception task, with the experimental condition and the participant group as independent variables and IQ and PPVT scores as covariates. Results showed a significant main effect of the participant group: children with ASD were less likely to point to the empty box than TD children in the deception tasks, $F(1,74)=12.97$, $p<0.001$. The main effect of the condition, $F(1,74)=0.53$, $p=0.47$, and the interaction between the group and the condition were not significant, $F(1,74)=0.37$, $p=0.54$.

Trial-by-trial Analysis

Similar to the distrust task, a trial-by-trial analysis was performed to test the group difference in the performance of the deception task for each trial. As shown in Fig. 4, in the social cue condition, the TD group performed significantly or marginally better than the ASD group in most of the trials. In the non-social cue condition, the TD group outperformed the ASD group in the first trial, $p<0.001$ (Fisher's Exact test, two-tailed), and 90% TD children placed the sticker on the empty box in the first trial. However, only 35% children with ASD succeeded in the first trial. The group difference attenuated in the following trials and disappeared in the last three trials, p's > 0.30 (Fisher's Exact test, two-tailed), as children with ASD learned quickly from the trial-by-trial feedback. In the social-cue deception condition, 71% TD children pointed to the empty box in the first trial, marginally more than the ASD group (38%), $p=0.086$ (Fisher's Exact test, two-tailed).

Survival Analysis

The proportions of participants who failed to learn the strategy at the end of each trial were computed for each group and each condition. As illustrated in Fig. 5, the survival analysis showed a similar trend as in the distrust task: in the social cue condition, 92.86% TD children learned to point to the empty box in the first two trials, while only 66.67% children with ASD learned at the end of the ten trials. In the non-social cue condition, when the social components of the tasks were reduced, 80.95% TD children and 66% children with ASD learned to point to the empty

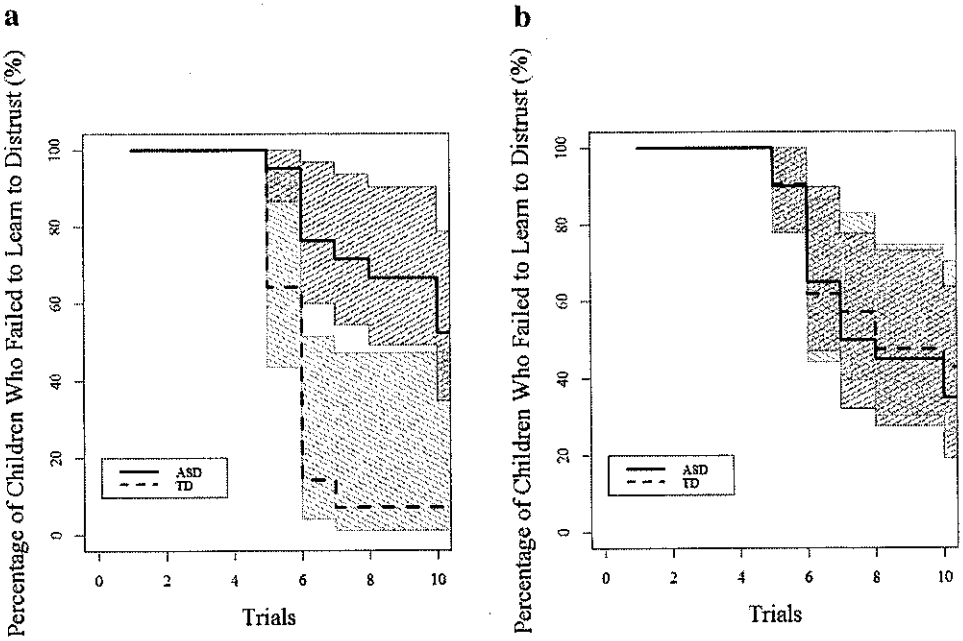


Fig. 3 Survival analyses. Percentages of children who failed to learn to choose the labeled box over trials in the social cue (**a**) and the non-social cue (**b**) conditions of the distrust tasks. The *shaded areas* represent the 95% confidence intervals

box at the end of the game. Log-rank tests showed that TD children learned faster than children with ASD in the social cue condition, $\chi^2(1)=13.4$, $p<0.001$, but the two groups had similar learning rates in the non-social cue condition, $\chi^2(1)=2.4$, $p=0.12$. When comparing the conditions within each group by the log-rank tests, we found that TD children learned faster in the social cue condition than in the non-social cue condition, $\chi^2(1)=3.9$, $p=0.049$, while children with ASD had similar learning rates in both conditions, $\chi^2(1)=0$, $p=0.97$.

Correlational Analyses

We also conducted correlational analyses to examine the possible associations between average scores of the distrust and deception tasks and the AQ, SRS, and SCQ scores in the ASD group. We did not find any significant correlation between the AQ, SRS, and SCQ scores and the trust and deception performance in ASD, p's > 0.05.

Discussion

The current study examines whether trust and deception behaviors in children with ASD are primarily affected by specific deficits in learning based on social cues. We recruited children with ASD and TD children aged from 4- to 8-years old, who had never been involved in our previous studies or any relevant research projects. Results of our social cue condition confirm the findings of Yi et al. (2014) regarding the atypical trust and deception behaviors in children with ASD: they were less likely and slower to learn to distrust and deceive the informant who repeatedly deceived them than TD children. Our novel findings came from the non-social cue condition when the social components of the distrust and deception tasks were controlled to reduce their impact: we observed a comparable learning ability in children with ASD relative to TD children. In other words, children with ASD had no difficulty learning these behaviors when the social cues were replaced by physical cues. These findings support the proposition that the abnormality of the trust and deception behaviors in children with ASD was mainly due to their defective learning in the social context (Bushwick 2001).

An alternative explanation of the specific difficulty in ASD children is that the social cue condition is cognitively more demanding simply because more information is available for processing. Compared to physical cues, the human actor brought in verbal instructions and gestural information in the tasks. This is a possible explanation for our results. The two tasks involve qualitatively similar cognitive processes, including general learning that cues are misleading based on trial-by-trial reward feedback, ignoring or inhibiting these cues, deciding to choose the alternative box and so on. The core difference is different ways to deliver the misleading information by using different cues (a person vs. a sticker). Interestingly, we found that TD children learned faster in the social cue conditions than in the non-social cue conditions despite extra available information. Thus, the presence of social cues facilitates rule understanding and problem solving in TD children, who are highly sensitive to social information. We speculate that this social strength or preference in TD children may facilitate their learning in the social context. In contrast,

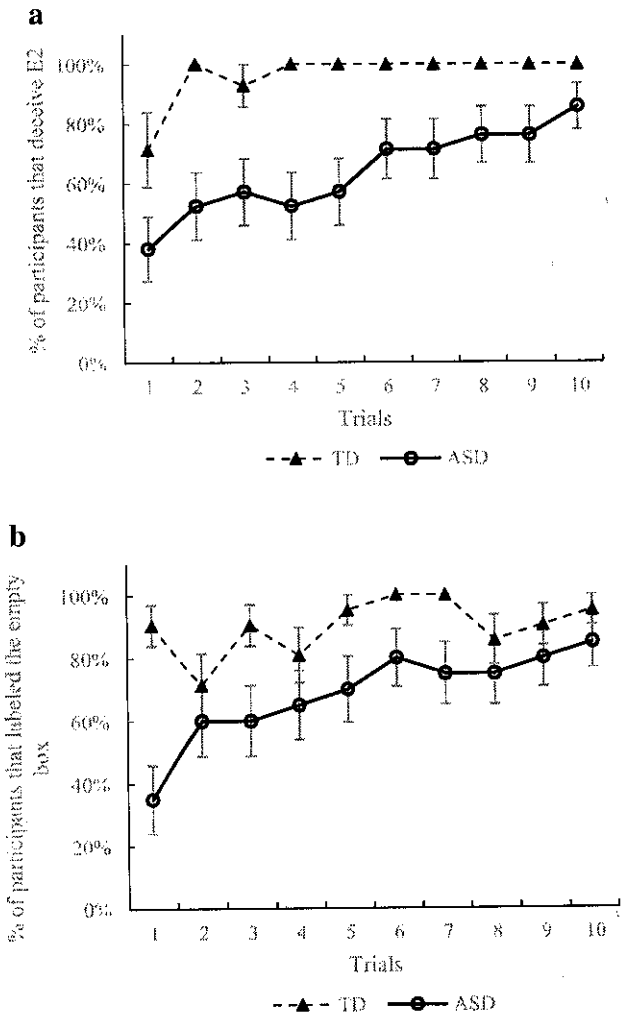


Fig. 4 Percentage of children with ASD and TD children who deceived E2 or labeled the empty box as a function of trial numbers in the social cue (**a**) and the non-social cue (**b**) conditions. *Error bars* represent 95% confidence intervals

children with ASD performed similarly in the social cue and the non-social cue conditions. Thus, the fact that they did not perform *worse* in the social cue condition suggests that extra cueing information (e.g., verbal instruction and gestural information) do not hamper the learning process. It is more likely that social strength or preference, manifested in the behaviors of TD children, is absent in children with ASD due to their diminished awareness and sensitivity to social cues. Future studies are warranted to test this hypothesis directly.

Another alternative way to explain our results of the non-social cue condition is that the two groups may interpret the non-social cues in different ways. Although there was no E2 in the non-social cue condition to interact with the child, the presence of E1 may make the TD child attribute social meanings to the physical marker as it was put by a person. At the same time, it is possible that children with ASD are less likely, compared to TD children, to attribute social meanings to physical cues in the non-social cue condition. In other words, children with ASD may interpret the physical cues in a mechanical way, while TD children may interpret it in a more social way. However, with behavioral paradigms we could not reverse engineer the exact strategies the two groups of participants used. Future research using brain imaging measures such as functional near-infrared spectroscopy (fNIRS) to explore the underlying mechanism of the tasks is therefore recommended.

Besides the findings above, we also discovered a carry-over effect for TD children in the deception task from the previous distrust task. As shown in Fig. 4, TD children in the deception task started with the better performance thus their learning curve was flatter than the distrust task. The carry-over effect by itself is an interesting observation. It might indicate some sort of learning generalization elicited by the previous distrust session, suggesting that social learning of TD children is flexible and generalizable to different conditions. However, such a generalization was not found in children with ASD in the same task, which confirms previous findings that the learning of children with ASD was less flexible and more constrained to specific circumstances relative to TD children (Yi et al. 2014).

Several limitations of the current study should be considered when interpreting our findings. First, limited by the characteristics of our tasks, the social components were not completely removed in our non-social cue condition. As discussed above, children may interpret physical markers in a social way. Despite this concern, removing the role of E2 has remarkably reduced the group differences in both distrust and deception tasks. This result suggests, in the social cue condition, that direct interaction with a person (E2) can cause significant difficulty for ASD to learn distrust and deception behaviors. Second, the cognitive load and complexity of these two conditions are not exactly the same. For example, in the social-cue condition E2 examined the boxes before the child made her judgment, while in the nonsocial cue condition this extra cue was absent. Future studies could introduce a mechanical device to attach the sticker before the child in order to make the available cues better matched between conditions. Third, we did not find any association between the trust and deception performance and the autism scales (AQ, SRS, and SCQ) in children with ASD. Future studies could use more specific social measures (e.g., pragmatic language, Theory of Mind) to explore a possible association between trust and deception and social abilities. Fourth, it is also noteworthy that our sample of ASD children is high-functioning, with an average IQ of approximately 100, which limits

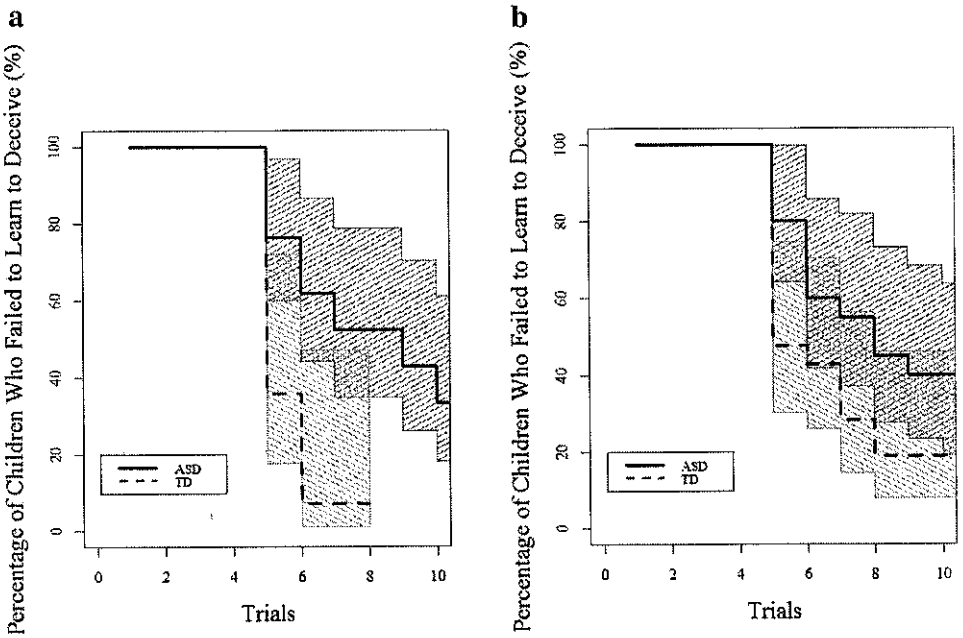


Fig. 5 Survival analyses. Percentages of children who failed to learn to deceive over trials in the social cue (**a**) and label the *empty box* in the non-social cue (**b**) conditions of the deception tasks. The *shaded areas* represent the 95% confidence intervals

the generalizability of these findings to lower-functioning children with ASD. Besides the social learning deficit, cognitive impairments may play a major role in both the social and non-social learning processes. Thus, the non-social learning could also be impaired in lower-functioning children with ASD. Our findings should also be interpreted with caution when extrapolated to children with ASD at different ages, as age plays an important role in children's development of trust and deception behaviors (e.g., Vanderbilt et al. 2011). Future research should include children with ASD with a broader age and IQ range, to investigate the developmental trajectory of social and non-social learning and the effect of cognitive dysfunction in this population. Fifth, our conclusions are based on a sample of Chinese children with ASD and TD children, so that cultural influences need to be taken into consideration when interpreting our findings. Generally speaking, Eastern and Western children exhibit highly similar distrust and deception behaviors (Fu et al. 2007; Xu et al. 2010), since both cultures generally encourage honesty and discourage deceit (Fu et al. 2010). However, subtle cultural variations of children's distrust and deception behaviors may exist, as a result of the emphasis on group harmony and collectivism in the East Asian culture (Rothbaum et al. 2000; Xu et al. 2010). For example, individualistic and collectivistic values influence the acceptability of lying, thus Chinese children prefer to lie to help the collective rather than the individual, while Canadian children do the reverse (e.g., Lau et al. 2013; Mealy et al. 2007; Sweet et al. 2010). Last, we used the between-subject design to minimize the learning generalization across conditions by randomly assigning the participants into the two conditions. With appropriate experimental designs, future research could compare one's performance in understanding the social vs. non-social cues using a within-subject design.

In summary, the current study aims to investigate whether reducing social cues could help ASD children overcome their deficits in trust and deception tasks. If so, this would provide evidence that their deficit in these tasks is more associated with their specific deficit in social learning than impairments in general learning ability. We found that the abnormalities in the distrust and deception behaviors in ASD are limited to the social situation, while their non-social learning is intact. Moreover, we found a social strength effect for TD children; their learning is benefited from the presence of social information. This social strength effect was absent in children with ASD due to their insensitivity to the social information. Besides, we also found a carry-over effect for TD children's non-social learning, which was absent in children with ASD. Overall, our findings support the social learning account of ASD (Bushwick 2001) in trust and deception behaviors. Our study has important implications for understanding the nature of ASD and interventions for individuals with the disorder. For example, when we train ASD children in social interactive tasks with similar complexity as our distrust and deception tasks, we should start with simple social cues in simple situations. After ASD children master the simple tasks, the training can incrementally include more social cues that are relatively challenging for them. Further work is needed to determine whether these findings could be generalized to a larger population of children with ASD.

Acknowledgments This work was supported by the National Natural Science Foundation of China (31571135, 31371020, 61533001, and 31622029), and the National High Technology Research and Development Program of China (863 Program, 2012AA011602). The authors are thankful to Dr. Lisa Joseph, Tianbi Li, Pengli, Li, Dr. Xueqin Wang, Qiandong Wang, Shuyuan Feng, Yinan Lv, and staff in Qingdao Elim School for their generous assistance in the project.

Author Contributions YY was responsible for data collection, data analysis, and manuscript preparation; YT was responsible for data analysis; JF was responsible for data collection; HL was responsible for data analysis and paper revision; KW was responsible for study design and manuscript preparation; LY was responsible for study design, data analysis, and manuscript preparation. All authors read and approved the manuscript.

Compliance with Ethical Standards

Conflict of interest All authors have no conflict of interest to declare.

Ethical Approval All procedures performed in this study were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Informed consent was obtained from all children included in the study and their parents.

References

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders, 4th text revision ed.* (pp. 553–557). Washington, DC: American Psychiatric Association.

Auyeung, B., Baron-Cohen, S., Wheelwright, S., & Allison, C. (2008). The autism spectrum quotient: Children's version (AQ-Child). *Journal of Autism and Developmental Disorders, 38*, 1230–1240. doi:10.1007/s10803-007-0504-z.

Baldwin, D. A. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology, 29*(5), 832.

Baron-Cohen, S. (1992). Out of sight or out of mind? Another look at deception in autism. *Journal of Child Psychology and Psychiatry, 33*(7), 1141–1155.

Bushwick, N. L. (2001). Social learning and the etiology of autism. *New Ideas in Psychology 19*, 49-75.

Chandler, M., Fritz, A. S., & Hala, S. (1989). Small-scale deceit: Deception as a marker of two-, three-, and four-year-olds' early theories of mind. *Child Development, 60*, 1263–1277.

Constantino, J. N., & Gruber, C. P. (2002). *The social responsiveness scale*. Los Angeles: Western Psychological Services.

Corriveau, K., & Harris, P. L. (2009). Choosing your informant: Weighing familiarity and recent accuracy. *Developmental Science, 12*(3), 426–437.

Dominey, P. F., & Dodane, C. (2004). Indeterminacy in language acquisition: The role of child directed speech and joint attention. *Journal of Neurolinguistics, 17*(2–3), 121–145.

Ewing, L., Caulfield, F., Read, A., & Rhodes, G. (2015). Appearance-based trust behavior is reduced in children with autism spectrum disorder. *Autism: The International Journal of Research and Practice, 19*(8), 1002–1009.

Fu, G., Brunet, M. K., Lv, Y., Ding, X., Heyman, G. D., Cameron, C. A., & Lee, K. (2010). Chinese children's moral evaluation of lies and truths—Roles of context and parental individualism-collectivism tendencies. *Infant and Child Development, 19*(5), 498–515.

Fu, G., Xu, F., Cameron, C. A., Leyman, G., & Lee, K. (2007). Cross-cultural differences in children's choices, categorizations, and evaluations of truths and lies. *Developmental Psychology, 43*(2), 278–293.

Hareli, S., Kafetsios, K., & Hess, U. (2015). A cross-cultural study on emotion expression and the learning of social norms. *Frontiers in Psychology, 6*, 1501. doi:10.3389/fpsyg.2015.01501.

Jones, E. J. H., Webb, S. J., Estes, A., & Dawson, G. (2013). Rule learning in autism: The role of reward type and social context. *Developmental Neuropsychology, 38*(1), 58–77.

Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development, 76*(6), 1261–1277.

Lau, Y. L., Cameron, C. A., Chieh, K. M., O'Leary, J., Fu, G., & Lee, K. (2013). Cultural differences in moral justifications enhance understanding of Chinese and Canadian children's moral decisions. *Journal of Cross-Cultural Psychology, 44*, 461–477.

Li, A. S., Kelly, E. A., Evans, A. D., & Lee, K. (2011). Exploring the ability to deceive in children with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 41*(2), 185–195.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. l., & Dilavore, P. C., ... Rutter, M. (2000). The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders, 30*(3), 205–223.

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 24*(5), 659–685.

Mealy, M., Stephan, W., & Urrutia, I. C. (2007). The acceptability of lies: A comparison of Ecuadorians and Euro-Americans. *International Journal of Intercultural Relations, 31*(6), 689–702.

Rothbaum, F., Pott, M., Azuma, H., Miyake, K., & Weisz, J. (2000). The development of close relationships in Japan and the United States: Paths of symbiotic harmony and generative tension. *Child Development, 71*(5), 1121–1142.

Russell, J., Mauthner, N., Sharpe, S., & Tidswell, T. (1991). The 'windows task' as a measure of strategic deception in preschoolers and autistic subjects. *British Journal of Developmental Psychology, 9*(2), 331–349.

Rutter, M., Bailey, A., & Lord, C. (2003). *The social communication questionnaire: Manual.* Torrance: Western Psychological Services.

Sodian, B., & Frith, U. (1992). Deception and sabotage in autistic, retarded and normal children. *Journal of Child Psychology and Psychiatry, 33*(3), 591–605.

Sodian, B., Taylor, C., & Harris, P. (1991). Early deception and the child's theory of mind: False trails and genuine markers. *Child Development, 62*, 468–483.

Sweet, M. A., Heyman, G. D., Fu, G., & Lee, K. (2010). Are there limits to collectivism? culture and children's reasoning about lying to conceal a group transgression. *Infant and Child Development, 19*(4), 422–442.

Talwar, V., Zwaigenbaum, L., Goulden, K. J., Manji, S., Loomes, C., & Rasmussen, C. (2012). Lie-telling behavior in children with autism and its relation to false-belief understanding. *Focus on Autism and Other Developmental Disabilities, 27*(2), 122–129.

Tomasello, M., & Farrar, M. J. (1986). Joint Attention and Early Language. *Child Development, 57*(6), 1454–1463.

Vanderbilt, K. E., Liu, D., & Heyman, G. D. (2011). The development of distrust. *Child Development, 82*(5), 1372–1380.

Wang, Z., Williamson, R. A., & Meltzoff, A. N. (2015). Imitation as a mechanism in cognitive development: A cross-cultural investigation of 4-year-old children's rule learning. *Cognitive Science, 6*, 562.

Xu, F., Bao, X., Fu, G., Talwar, V., & Lee, K. (2010). Lying and truth-telling in children: From conception to action. *Child Development, 81*(2), 581–596.

Yi, L., Fan, Y., Li, J., Huang, D., Wang, X., Tan, W., … Lee, K. (2014). Distrust and retaliatory deception in children with Autism Spectrum Disorder. *Research in Autism Spectrum Disorders, 8*(12), 1741–1755.

Yi, L., Pan, J., Fan, Y., Zou, X., Wang, X., & Lee, K. (2013). Children with autism spectrum disorder are more trusting than typically developing children. *Journal of Experimental Child Psychology, 116*(3), 755–761.

Yirmiya, N., Solomonica-Levi, D., & Shulman, C. (1996). The ability to manipulate behavior and to understand manipulation of beliefs: A comparison of individuals with autism, mental retardation, and normal development. *Developmental Psychology, 32*(1), 62.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing, 70*(13–15), 2149–2165.

Philosophy

DOI: 10.1111/ejop.12099

The Limits of Learning: Habermas' Social Theory and Religion

Maeve Cooke

Abstract: Habermas' view that contemporary philosophy and social theory can learn from religious traditions calls for closer consideration. He is correct to hold that religious traditions constitute a reservoir of potentially important meanings that can be critically appropriated without emptying them of their motivating and inspirational power. However, contrary to what he implies, his theory allows for learning from religion only to a very limited degree. This is due to two core elements of his conceptual framework, both of which are key features of his account of postmetaphysical thinking. The first is the requirement of ethical agnosticism; this requires philosophy and social theory to refrain from offering guidance on questions of the good life. The second is his language-immanent conception of truth in the domain of practical reason; this follows from his rejection of any source of validity beyond human communication in this domain. I make the case for a more robust account of learning from religious traditions and metaphysical worldviews, arguing that for this purpose Habermas must modify his requirement of ethical agnosticism and relinquish his language-immanent conception of truth.

Social Learning Theory

Learning—learning from history and learning from other cultures—has long been a central motif in Habermas' critical social theory (Habermas 2001: 38–57; Habermas 1992: 138). In recent years, he has begun to underscore the importance of learning from religion: philosophy and social theory are encouraged to learn from religious traditions and religious believers and non-believers are urged to engage in processes of mutual learning. His new emphasis on learning from religion may come as a surprise to those familiar with his writings of the 1980s and 1990s, in which his attitude to religion is, at best, distanced. To all appearances, at least, *The Theory of Communicative Action*, first published in 1981, subscribes to the Weberian view that religious belief will die away as modernity progresses. While in his subsequent writings over the next two decades he occasionally suggests that religion might continue to have a place within modernity, the tone of his remarks is hesitant and the language he uses is cautious. In light of this it is interesting to see that the theme of learning from religion is already present in his writings of the 1980s. Thus in an essay first published in 1988, Habermas calls on postmetaphysical thinking to appropriate critically the 'old truths' of metaphysics in general, and of religion in particular (Habermas 1992: 14–5). Evidently, therefore, postmetaphysical philosophy and social theory have something to learn from religion. But learning requires what

European Journal of Philosophy 24:3 ISSN 0966-8373 pp. 694–711

Habermas here calls 'critical appropriation' (in later texts he speaks of 'translation'). The need for such appropriation is attributed to the modern plurality of worldviews: the inhabitants of modernity lack a shared evaluative basis that would permit the truths of metaphysics and religion to be accepted by everyone on the basis of good reasons. This idea of critical appropriation anticipates his more recent reflections on learning from religion, in which he emphasizes religion as a reservoir of meaning that can serve as a semantic resource for postmetaphysical philosophy and social theory, as well as for the citizens of modern secular states (Habermas 2008). As before, the aim of critical appropriation is to salvage the truths of religious traditions and recast them in a generally accessible language so that everyone could accept them on the basis of good reasons, irrespective of particular religious beliefs or conceptions of the good.

Habermas' claim that religious traditions provide a valuable semantic resource for postmetaphysical philosophy and social theory calls for closer consideration. Before examining it, let me make clear that I agree both with his view that religious traditions constitute a reservoir of potentially important meanings and with his view that these meanings can be critically appropriated—translated—without emptying them of their motivating and inspirational power. In short, I agree that contemporary philosophy and social theory may learn from religion by way of (what Habermas now calls) translation. I want to argue, however, that in Habermas' critical social theory the possibility of learning from religion is highly restricted. Indeed, it is much more restricted than his recent writings give us to understand (Habermas 2008, 2006, 2010). This is due to core elements of his conceptual framework. I see two conceptual obstacles preventing him from developing a more robust conception of learning from religion. Both are key features of his idea of postmetaphysical thinking. The first is his insistence that postmetaphysical thinking must abstain from offering guidance with regard to questions of the good life and good society. In the essay referred to above, when distinguishing postmetaphysical from metaphysical thinking, he initially identifies the core issue as the 'enlightening role' of philosophy with regard to life practices as a whole.[1] Whereas metaphysical thinking offers orientation in matters relating to the conduct of life and to human flourishing in general, postmetaphysical thinking is abstinent in this regard (Habermas 1992, 14–5). I refer to this as the thesis of ethical agnosticism. The second is his insistence that in the domain of practical reason postmetaphysical thinking may not affirm any notion of validity that has a source beyond the human world of linguistic communication (Habermas 2003b: 11–5). This is part of his thesis of 'innerworldly transcendence' or 'transcendence from within' (Habermas 1996: 5, 17; Habermas 2003b: 10–1). He proposes a deflationary interpretation of the 'unconditioned' or 'absolute', according to which the transcending power of reason is contained *within* the forms of communication through which human beings reach an understanding with one another. As he puts it, 'the transcendental tension between the ideal and the real, between the realm of the intelligible and the realm of appearances, enters into the social reality of situated interactions and institutions' (Habermas 2008: 25). I refer to this as his language-

immanent conception of practical validity or truth (I have more to say about what I mean by truth below). I argue that his ethical agnosticism leads to an impoverished conception of learning from religion; moreover, one that sits uneasily with his remarks on the importance of the critical appropriation of the truths of religious traditions. I then show that a more adequate account of learning from religion would require him to modify his sharp distinction between morality and ethics and, in the domain of practical reason, to give up his language-immanent conception of validity.

Let us look more closely at what Habermas means by critical appropriation of the 'old truths' of metaphysics and religion. As already mentioned, this theme is to the fore in his recent writings, where he emphasizes the importance of religion as a semantic resource for postmetaphysical philosophy and social theory (and also for the citizens of secular states, although I leave this question aside in the following[2]). Thus, in an essay on Kant, he describes religious traditions as possible reservoirs of meaning that are capable of exercising an inspirational force on society *as a whole* (Habermas 2008: 142). He criticizes Kant for according a purely instrumental function to positive religion and ecclesiastical faith, proposing instead that religious traditions, with their striking models, their vivid exemplary figures, their inspirational stories of the lives of the saints and prophets, their promises, their miracles, their suggestive images and edifying narratives, can stimulate the imagination of secular as well as religious citizens, motivating them, for instance, to work collectively towards realizing on earth a secular version of the promised kingdom of God (ibid: 223–47). As he sees it, secular translations of religious projections of successful forms of life can continue to inspire us, and encourage us to make tentative efforts at cooperation with a view to bringing about social change for the better, even without the certainty of divine assistance (ibid: 227; Habermas 2010: 18–9). For this purpose we require a form of critical appropriation or, as he now tends to say, translation, that salvages the substance of a term without deflating or exhausting it (Habermas 2006: 45). Examples include the translation of the concept of 'man [made] in the image of God' into the idea that the equal dignity of human beings deserves unconditional respect (ibid), Walter Benjamin's idea of anamnetic solidarity with past injustice, which Habermas describes as a concept 'manifestly trying to fill the gap left by the lost hope in a Last Judgment' (Habermas 2008: 241) and Marx's idea of the emancipated society, which he sees as a secular version of the kingdom of God (ibid: 231). Recall that translations recast the contents of religion in a generally accessible language that would permit everyone to accept them on the basis of good reasons, irrespective of religious belief and irrespective of their particular conceptions of the good. Recall, too, that the point of translation is semantic regeneration. By this I take Habermas to mean the provision of meanings that human beings can draw on in their various life-practices. Such meanings help us to make sense of the situations in which we find ourselves, they give us reasons for acting and they inspire us to explore new paths in life. In other words, they offer us ethical orientation: they position us in space and time, point us in certain directions and open up new possibilities.

However, as we have also seen, postmetaphysical thinking is supposed to be ethically agnostic: it is supposed to abstain from offering guidance in matters relating to the good life and good society. Moreover, it is supposed to articulate its views so that they are in principle *rationally* acceptable to everyone. How are we to make sense of what appear to be two conflicting demands? On the one hand, translation is required for the purposes of ethical orientation; on the other hand, postmetaphysical thinking is required to refrain from offering ethical orientation. How can Habermas deal with this apparent conflict? How can he allow for semantic regeneration by way of translation without compromising the ethical agnosticism of postmetaphysical thinking and the need for the translated contents to be acceptable to everyone on the basis of good reasons?

Habermas himself does not mention this difficulty, much less propose a solution to it. This may be because he believes he can avoid it with the help of his distinction between ethics and morality. And, in fact, this *is* one way of dealing with it. The conflict can be avoided by restricting the learning process to what he calls strictly moral matters, understood as matters relating to the just regulation of interpersonal conflicts from the point of view of the universalizability of interests. Postmetaphysical thinking can draw on the reservoir of meanings offered by religious traditions without violating the requirement of ethical agnosticism if it limits the possibility of learning from religion to questions of justice, leaving the domain of ethics completely to one side.

As is well known, in Habermas' conception of practical reason, moral questions are distinguished sharply from ethical ones (Habermas 1993). Ethics is concerned with questions of individual and collective self-realization, which are raised by individuals and groups in specific, local contexts; it deals with questions of the good life and good society. By contrast, morality is concerned with self-determination; it deals with matters of justice, by which he means it serves to regulate interpersonal conflicts through reference to the principle of universalizability of interests. Postmetaphysical thinking does not preempt the answers to moral questions. Nonetheless, it provides a framework for deliberation on such questions in the form of an account of communicative rationality; when applied specifically to moral questions, communicative rationality yields a formal moral theory.[3] This moral theory offers a framework for reflection on moral matters that permits them to be assessed from the point of view of a truth-analogous conception of validity, which, as such, is binding on everyone, irrespective of historical and social context.

Unlike moral questions, which raise claims to validity in a truth-analogous sense, ethical questions are held to be context-specific, not just in the sense that they emerge as questions only in specific historical and social contexts, but also in the sense that they admit of answers that hold only for particular individuals and groups and are thus not universally binding. In line with this, ethical discourses are described as intersubjective deliberations concerned with the hermeneutic explication of individual and collective value orientations and self-understandings, as opposed to seeking an outcome that is universally valid. In contrast to its contribution in the domain of morality, postmetaphysical

thinking provides no framework for thinking about the validity of ethical questions: it leaves the question of ethical truth to one side. If we take seriously Habermas' distinction between morality and ethics, therefore, his call for critical appropriation of the truths of religion must be taken to refer only to moral matters; he must envisage these truths as feeding into processes of *moral* as opposed to ethical deliberation.

Some of his examples clearly support this reading of his position. Think of his example of the translation of the religious view that human beings are made in God's image into the postmetaphysical view that human beings have an inherent equal dignity, entailing the requirement of equal respect. Postmetaphysical moral theory builds this insight into its account of moral deliberation in the form of a discourse theory in which all persons affected are entitled to participate fully in the deliberation and to have their interests equally taken into account (Habermas 1990).[4] On the level of postmetaphysical discussion of concrete moral questions, this insight guides the deliberative process in a tangible way, calling on participants in discussion publicly to justify any apparent inequalities relating either to the conduct of the deliberation or to its outcome.

This reading is further supported by remarks he makes in an essay discussing what he describes as Kierkegaard's postmetaphysical approach to ethics. He writes that Kierkegaard's 'postmetaphysical abstention runs up against its limits in an interesting way as soon as questions of a 'species ethics' arise. As soon as the ethical self-understanding of language-using agents is at stake *in its entirety*, philosophy can no longer avoid taking a substantive position' (Habermas 2003b: 14–5). Habermas holds that this is the situation in which we currently find ourselves due to the advance of the biological sciences and biotechnological developments. He continues: '[these developments raise] moral questions of an *altogether different kind;* questions that touch on the ethical self-understanding of humanity as a whole'. He then observes that such questions take us *beyond* postmetaphysical thinking. This clearly points towards the reading of his position I have offered in the foregoing, according to which postmetaphysical thinking leaves questions of ethical validity to one side. At the same time, however, it shows the need for an alternative conception of postmetaphysical thinking as it relates to ethics. Habermas acknowledges this implicitly for, in an essay in the same volume, he intervenes directly in the debate over genetic engineering.[5] He justifies his intervention by expressing his concern for how biotechnological developments may deeply affect our ethical self-understanding as members of the human species, with far-reaching consequences for our self-understanding as moral beings. In my view he is right to intervene, moreover to do so as a philosopher and social theorist (as opposed simply to a concerned citizen or public intellectual). For, as I see it, philosophy and social theory have valuable contributions to make to reflection on, and discussion of, matters of such profound importance for human flourishing. If questions of human flourishing and the conduct of life were to be removed from the realm of philosophy and social theory, the result would be a very restricted conception of their tasks. We might add that seeing such questions as beyond the remit of

philosophy and social theory is at odds with the self-understanding of earlier critical theorists in the Frankfurt School tradition, from which Habermas' theory has emerged; these theorists consistently emphasized critical theory's concern with human emancipation. Max Horkheimer famously described critical theories as seeking to 'liberate human beings from the circumstances that enslave them' (Horkheimer 1982: 246).[6] Evidently, Horkheimer is here advocating a mode of theorizing that would help us to attain a social condition in which we could flourish as human beings.

I have suggested that Habermas' insistence on the need for ethical agnosticism leads to a restricted conception of learning from religion and also to an impoverished conception of critical social theory. In addition, using the example of genetic engineering, I pointed out that he proves unable to maintain the ethically agnostic stance he himself advocates. His difficulty in consistently upholding the thesis of ethical agnosticism is evident, too, in the examples he gives of successful translation. For while some of these, as mentioned, clearly relate to moral matters, others seem to belong to the domain of ethics. Think, for instance, of Marx's translation into secular terms of the religious idea of the Kingdom of God; the social condition projected by Marx is one in which human beings have been emancipated from their servitude and are now free to lead a 'good life' in which they develop their particular capacities and, in general, flourish as human beings. Indeed, the overall tone of Habermas' recent writings on religion, in which he places emphasis on regeneration of the semantic fabric of society *as a whole*, suggests that what he has in mind is not just *moral* regeneration, in the narrow sense of providing resources for regulating interpersonal conflicts, but also ethical regeneration, in the sense of providing resources for life-practices that foster human flourishing.

The obvious conclusion to draw from the foregoing is that Habermas should relinquish or modify the requirement of ethical agnosticism and open up philosophy and social theory to the possibility of learning from religion in ethical as well as moral matters. But this is not an easy option for Habermas. Were he to do so, he would have to make fundamental changes to the conceptual framework of his theory. For one thing, he would have to relax his sharp distinction between morality and ethics, which is the basis for his justification of a universalist moral theory.[7] For another, he would have to give up his language-immanent conception of (practical) truth, since, as we shall see, it is incompatible with learning from religion in ethical matters.

The discussion so far has helped to clarify what it might mean for philosophy and social theory to *learn* from religion. There are, of course, different kinds of learning.[8] One kind of learning is aimed primarily at the acquisition of certain techniques or skills. But, clearly, this is not what Habermas has in mind in the present instance. A second kind of learning is aimed primarily at improving self-knowledge on the part of individuals or groups. This appears to be part of what Habermas has in mind when he refers to the need for mutual learning processes on the part of both religious and non-religious citizens (Habermas 2008). Learning in this case seems to mean a self-reflexive attitude regarding one's own

view of the world, be it secular or religious. For non-believers it means, for example, self-conscious awareness that secular thinking has its historical roots in religious beliefs and traditions, that religion historically has been a force for world-disclosure and semantic renewal, that it has at times also been a force for progressive social change, and that today it retains an inspirational force for some people. For religious believers it means coming to terms with the fact of religious pluralism, with the emergence of modern science and with the spread of positive law and secular morality; it could also mean self-conscious awareness of the entwinement of religion with violence, domination and repression, though Habermas does not emphasize this especially. But this is not necessarily *mutual* learning. Indeed, in Habermas' account, the envisaged work of self-reflection on the part of secular and religious citizens seems to take place independently of one another: what he describes is not mutual learning but two complementary learning processes that run on separate tracks.[9]

The kind of learning involved in critical appropriation of the 'old truths' of religion (and of metaphysics in general) is different to this. When Habermas calls for critical appropriation or translation he apparently envisages not just improved self-knowledge but, over and above this, a kind of learning akin to a Gadamerian fusion of horizons. The latter is his model for intercultural learning. Drawing on Hans-Georg Gadamer's idea of the fusion of interpretative horizons, he sees intercultural understanding not as an assimilation, but as a convergence of perspectives, in which it is necessary for each of the culturally diverging parties to attempt to grasp things from the perspective of the other (Gadamer 1979). In this conception even profound cultural disagreements are contexts of possible learning. As Habermas makes clear, the conception rests on the presupposition that competing forms of life share an orientation towards culture-transcending ideas of truth and moral rightness; it is these culture-transcending, universalist ideas that allow individuals and groups with opposing conceptions of the good to enter into unrestricted dialogue with one another with a view to reaching agreement, instead of persisting in their claims to exclusive validity in a fundamentalist way.

Habermas does not describe learning from religion as a form of intercultural learning. Nonetheless, his conception of critical appropriation fits well with this model. Like intercultural learning, it is a process of *engagement*, in which the quest for understanding is coupled with an attitude of critical evaluation. Recall, however, that in the intercultural case, such engagement is possible only because the process of critical evaluation is guided by ideas of truth and moral rightness that are culture-transcending. If, as I have proposed, learning from religion is to include ethical as well as moral matters, Habermas will have to supply a culture-transcending idea of practical validity that is not restricted to moral questions in the narrow sense but broad enough to encompass ethical ones as well.

As things stand, this is not an option for Habermas. He himself is reluctant to move in this direction, citing a concern to respect the modern plurality of worldviews. As he sees it, the inhabitants of modernity lack a common framework that would enable them to reach agreement on substantive matters of the

good life and good society no matter how long they discussed these matters and no matter how satisfactory the argumentative conditions under which they did so. In my view, however, the difficulty is not just the *factual* plurality of worldviews in modernity and the *unlikelihood* that it will give way to ethical homogeneity; rather it is a deeper, conceptual one. The difficulty is that his language-immanent account of practical validity cannot accommodate the experientially based, disclosive moment of ethical validity claims. When it comes to questions of the good life and good society, in many cases we will fail to be convinced by the arguments of others until we have undergone an epistemically (and existentially) significant shift in perception. Such shifts in perception involve truth as disclosure, whereby the world, or some aspect of it, appears in a different light and we gain a new way of looking at things. While the required shift in perception may be brought about argumentatively, as when we are swayed by the arguments of others, it typically depends on non-argumentative experiences, which make us receptive to these arguments or, indeed, substitute for them.[10] As I aim to show, truth as experientially based disclosure finds no place within the framework of postmetaphysical thinking, with its thesis of innerworldly transcendence and corresponding language-immanent conception of practical validity (truth).

This difficulty arises in the domain of practical reason in general but is especially relevant to our present discussion. For, as we have seen, learning from religion calls for critical appropriation of ethical contents that stimulate the imagination and are inspirational.[11] I want to suggest that their inspirational force depends on truth as disclosure.[12]

In order to give a sense of what I mean by truth as disclosure, I consider what it might mean critically to appropriate the meaning of an exemplary figure or action. Adopting Habermas' recent terminology, I treat this as a question about translation.[13] Our question is whether stories of exemplary figures and actions can be translated in such a way that their exemplary validity is potentially rationally acceptable to the inhabitants of modernity.[14] It seems clear that in such cases, translation is not simply a matter of rephrasing or reformulating propositional contents in a different idiom. My contention is that it should be understood as a matter of creatively rearticulating the inspirational contents in a way that produces subjective experiences in which truth is manifested anew. My hunch is that an adequate account would show that a successful translation of this kind does at least three things. It re-presents the truth of the original, it resonates with the subjectivity of its addressees and it opens their eyes to new ways of seeing themselves and the world. Translation of this kind is an aesthetic activity in the Kantian sense of the term: it involves opening up new spaces of the imagination in which truth can appear in a new time. The space it opens is one in which by way of new interpretations of self and world, and by way of a new interplay of linguistic expressions and modes of articulation, truth is re-presented: it appears to us anew.

I have chosen to focus on the translation of stories of exemplary figures and actions because critical appropriation seems particularly difficult in such cases.[15]

Indeed, we might wonder whether it is at all possible to translate stories of exemplary figures and actions in a way that does not deflate them or empty them of their truth content. Take, for instance, the biblical story of Abraham's willingness unconditionally to obey God's command to sacrifice his beloved son, Isaac.[16] On some readings of this story, Abraham is intended as an exemplary figure.[17] But even for religious believers in the Judaeo-Christian tradition, it is not at all clear why one should regard him as such; how much harder, then, for non-believers to feel the power of the example he allegedly sets. The difficulty, in my view, is that the meaning of exemplary figures and acts has an irreducible experiential dimension that is contextually and subjectively mediated. In other words, understanding exemplarity depends on a subjective response that has an affective as well as cognitive element and is, in turn, dependent on multiple factors relating to the historically specific, socio-cultural context in which the subjects addressed find themselves, as well as to the biography and psychic constitution of these subjects. This experiential component significantly complicates the translation exercise. However, as we shall see, it renders it neither impossible nor unproductive.

To speak of exemplarity is to speak of the normativity of the example: its power to inspire and to motivate us in our thinking and action. There have been some recent attempts among philosophers to offer an account of exemplary validity in terms of the Kantian paradigm of reflective judgment (e.g. Ferrara 2008). What I find missing from such accounts is consideration of the mode of truth involved in such judgment. Drawing on the work of Paul Ricoeur, I want to suggest that it is truth as disclosure: the example imaginatively opens up a space in which truth appears (Ricoeur 1995).[18] At the same time, again drawing on Ricoeur, I want to emphasize that the exemplary power of figures and actions is always mediated by contexts of interpretation. Even the most seemingly immediate experience of truth is mediated by the subjectivity of the person experiencing it. This subjectivity is in turn influenced by the historical and socio-cultural context inhabited by the person in question, and by the linguistic practices of that context. We could say: truth is always mediated truth, even when its operative mode is disclosure: even disclosed truth is *articulated* truth (ibid: 48–67).

Given the mediation of disclosed truth by contexts of interpretation and its dependence on experiential factors of a subjective and contextual kind, how is it possible critically to appropriate stories of exemplary figures and acts without emptying them of their meaning and robbing them of their motivating and inspirational power? Let us consider the exemplary act presented in the following story.

The story is taken from Orhan Pamuk's novel, *My Name is Red*, which is set in Istanbul in the 1590s (Pamuk 2001). The practice of manuscript illumination is at the centre of the book. One of the main characters is Master Osman, the master illuminator in charge of the workshop in which miniaturists are trained to illuminate manuscripts for the Sultan's palace in Istanbul. Master Osman is inspired by stories of the great master illuminators who have preceded him in

the tradition, in particular by the master illuminators who produce their best work when blind, guiding their brushes from memories acquired after years of thought, contemplation and reflection. Some of these masters went blind as a result of the fervour of their work, but some of them are supposed to have blinded themselves voluntarily (ibid: 393). Part of the plot of the novel entails Master Osman being granted access to the Sultan's otherwise locked and closely guarded treasury, where he is free to examine the most precious manuscripts of the greatest master illuminators. One of these is the Great Master Bihzad, who was reputedly blind in the last years of his life. Master Osman is convinced that the Great Master Bihzad blinded himself deliberately (ibid: 391). When examining one of the manuscripts in the treasury, Master Osman finds a reference to the 'turquoise- and mother-of-pearl-handled golden plume needle which [. . .] Master of Master Illuminators, Bihzad, used in the act of blinding his exalted self' (ibid: 392). Master Osman finds that very needle in the treasury and uses it to blind himself, smiling as he presses the needle into the pupil of first his right eye and then his left (ibid: 393–4).

The Great Master Bihzad's act of blinding himself inspires Master Osman and motivates him to perform *his* act of self-blinding. Clearly, it offers him ethical orientation. We might also say that Bihzad's act of self-blinding has *exemplary validity* for Master Osman. The crucial question, however, is whether it can be translated in such a way that it has exemplary validity *for us*, the inhabitants of a quite different cultural context, with quite different ideas of what makes an action ethically desirable or repugnant.

I have emphasized that truth as disclosure is always mediated truth. To feel the full force of the example in the Bihzad-Osman case, we would have to inhabit a complex interpretative framework in which the purpose of illumination is to evoke and honour the divine, in which the hand of the master illuminator is merely a vehicle for the divine, in which blindness is a gift of God bestowed on those masters who, by arduous discipline and training, come so close to the divine that they no longer need physical eyesight for their works of illumination, and so on. Thus, it is nearly impossible for us, the inhabitants of a quite different historical and socio-cultural context, to feel the full force of the example in this case. Nonetheless, with the help of a translation, we may be able to feel its power in some measure, to a greater or lesser degree; moreover, by feeling its power in some measure, we gain initial access to a context of meaning that was previously alien to us; this may enable us eventually to grasp it and, perhaps, value it more deeply. Two observations may be helpful here.

First, contexts of interpretation, expression and articulation are rarely so foreign to us that we are unable to gain any imaginative access to them whatsoever. That is to say: truth's mediations are rarely completely unfamiliar to us; some points of connection are almost always available. In the Bihzad-Osman case, we may be able to gain imaginative access to the space in which truth appears by connecting, for instance, with the idea that vision is not just a matter of eyesight but a function of bodily memory as a whole and/or with the idea that the love of beauty can inspire us to act courageously and/or the idea that

the excellence we gain through self-discipline and training brings us to a higher level of consciousness, perhaps even that it brings us closer to God. In other words, the force of the example can be experienced more or less immediately, but is rarely completely inaccessible to us.

Second, a powerful—poetically gifted—writer/translator can enable us to feel the normative force of exemplary acts even in contexts that are not readily accessible to us. It testifies to Pamuk's poetic talent as a writer of stories that his readers are likely to find Master Osman's emulation of Great Master Bihzad's act of self-blinding powerful; Pamuk invites us into an interpretative context that is alien to us in many respects; yet he succeeds in making us feel sufficiently at home there to be able to experience the exemplarity of the figure of Bihzad, not with full force but at least to some degree. In using the word 'powerful' I do not want to give the impression that I see the writer/translator as determining or controlling the transfer of meaning; rather, the genius of the writer/translator consists in her or his role as facilitator. Indeed, I wish to emphasize that the activity of critical appropriation depends crucially on the subjective activity of the readers or, more generally, receiving subjects. In other words, the writer as translator facilitates the appropriation of contexts of meaning by his or her receiving subjects in a process that involves active engagement of their subjectivity.[19]

But why should we speak of truth at all in this context? It could be objected that I have simply assumed that the power of the example to inspire and motivate is connected with a culture-transcending, truth-analogous claim to validity. This objection is warranted. I have provided no argument in support of my claim that truth appears in exemplary figures and actions. This is mainly because I have not yet worked out the requisite argument to my satisfaction. Until I do so, I can merely point to its guiding intuition. The intuition is that we could not adequately account for the inspirational and motivating power of ethical and religious experiences unless we assumed that something intrinsically valuable was disclosed to the subjects of these experiences. Nor could we adequately account for the transformative aspect of the disclosure—its power to bring about an existentially and epistemically significant shift in perception—unless we assumed that something intrinsically valuable was involved. 'Truth' seems an appropriate name for the intrinsically valuable substance of disclosures of this kind. Furthermore, in a Habermasian vein, I maintain that we would be unable to engage critically with the exemplarity of the figure or action without situating it in a cultural context with whose inhabitants we share a concern for validity in a culture-transcending, truth-analogous sense.

I wish to stress the epistemic unreliability of disclosures of this kind. This is due to the mediated nature of all experience: a subjective experience alone does not amount to reliable *knowledge* of the validity of what is disclosed. The subjects in question make claims on behalf of the intrinsic value of what appears to them, which (subject to contextual considerations) have to be opened to critical scrutiny in public processes of argumentative evaluation.[20] I want to emphasize, too, that truth as I conceive it is *inherently* culture-transcending: it constitutes a reference point for inquiry that always escapes our attempts to capture it; thus,

complete knowledge of truth is not available to human beings.[21] The thesis of the culture-transcending nature of truth, coupled with the thesis that truth is disclosed in experiences of exemplary figures and actions, means that the *religious* dimension of the story of Bihzad/Osman (or, indeed, Abraham) is not crucial for the purposes of the present discussion. In other words, if we experience these figures as exemplary, it is because something of intrinsic value appears to us when we encounter them. While making sense of its intrinsic value involves a culture-transcending idea of validity, this idea of validity is not specific to any religion and shares only certain features with traditional religious conceptions of the divine.

Do translations of inspirational figures and actions offer ethical orientation in a way that respects the modern plurality of worldviews? Whereas Habermas takes respect for the plurality of worldviews to require ethical agnosticism, such agnosticism, as we have seen, unacceptably limits the possibilities of learning from metaphysical and religious traditions and is out of tune with Habermas' own emphasis on the importance of semantic regeneration. This is why I argue that he should give up his ethically agnostic position. At the same time, I share his concern to respect the plurality of worldviews. But it seems to me that such respect requires not ethical agnosticism, but a normative commitment to individual autonomy and rational accountability—two normative principles at the heart of Habermas' theory of communicative action. In other words, respecting the plurality of worldviews does not prohibit philosophy and social theory from translating the contents of metaphysical and religious traditions for the purposes of ethical orientation: it merely requires them to acknowledge that those they address must work out for themselves, on the basis of good reasons, the ethical validity of the translated contents. In my view Pamuk's translation respects the plurality of worldviews in the sense of upholding the normative principles of individual autonomy and rational accountability. An in-depth discussion of the story, in particular of the techniques it employs to provoke reflexivity in the reader, would show that an aspect of the good life (in this case, self-blinding) is not imposed on those who hear or read it;[22] rather it is left up to those it addresses to accept or reject the validity of Master Osman's action, moreover to do so for reasons of their own that they could attempt to justify to others if called upon to do so. I emphasized that *whether* we feel the force of the example is contingent on a host of complex factors. But if we do feel its force, at least to some degree, it will not be because we have been compelled unquestioningly to accept the truths of the metaphysical tradition to which both Bihzad and Osman belong; rather, it will be because Pamuk has helped us to gain imaginative access to this metaphysical tradition and enabled us to make connections between what we find there and our own experiences, actions and judgments. In other words, Pamuk will have enabled the *critical* appropriation by the inhabitants of modernity of an inspirational and motivating content.

All this suggests that Habermas does not need the thesis of ethical agnosticism, understood as the requirement that philosophy and social theory abstain from offering guidance in matters of the good life and good society. Since

relinquishing this version of it would enrich his account of learning from religion (and metaphysics in general), I take the view that he should give it up. However, as mentioned, an enriched account, which would allow for ethical learning from religion, would have profound implications for the conceptual framework of his theory. As already indicated, it would require him to modify his strict distinction between morality and ethics, which would have significant implications for his account of moral validity in terms of the universalizability of interests. In addition, it would require him to give up his language-immanent conception of practical validity. I will conclude by outlining why this would be necessary.

Habermas' language-immanent conception of practical validity is most clearly evident in his discourse ethics. His rejection of anything outside of, or beyond, human practices of linguistic communication leads him to propose what we may call an *epistemic-constructivist* conception of moral truth. In his view, moral reality—the moral world—is constructed. It is 'made by us' (Habermas 2003a: 266–75). The realm of morality is itself generated in discourse. As he puts it: in the case of moral validity, the very domain of validity has to be *produced* (ibid: 262).[23] Moral validity is defined as the outcome of an idealized discursive procedure, universal in extent, in which participants reach agreement that a particular norm or principle is equally in everyone's interests. This discursively generated agreement is conceived of as the 'single right answer' to the question at hand; moreover, as an answer that is arrived at by everyone for the same reasons. This is what lends moral validity its unconditional, universally binding character and makes it analogous to truth (Habermas 2003a: 247–9).

I have a number of objections to this epistemic-constructivist account of practical validity. For one thing, as already indicated, it is unable to accommodate the experientially based, disclosive moment of practical reasoning. As I have argued in more detail elsewhere, in the domain of practical reason the rational acceptability of validity claims may depend on shifts in perception that are not attributable primarily to the exchange of arguments (Cooke 2013a; Cooke 2014). Think, for example, of trying to convince someone by reasoned argument that she should become a vegetarian. In many cases no amount of reasoning will work until the person in question has come to see the world in a way that make her receptive to the reasons for being a vegetarian (perhaps as a result of an existentially significant experience such as a visit to an abattoir). But as soon as we allow for subjective experiences in specific contexts as possible co-determinants of rational acceptability, the Habermasian conditions for generating unconditionally valid knowledge (truth) no longer obtain. No matter how perfect the argumentative conditions, in any given instance the rational acceptability of a particular argument might depend on argumentation-external factors absent in that instance. This would mean in turn that even perfect argumentative conditions would not guarantee a consensus and, hence, that an argumentatively achieved consensus could not *produce* truth. In other words, if the experientially based, disclosive moment of practical reasoning is given its due, there can be no guarantee that

participants in deliberation would agree on the rational acceptability of moral norms or principles *even under ideal discursive conditions*.

In addition, I have some more general reservations. In a nutshell, my general objections are that Habermas' epistemic-constructivist account of moral validity is hubristic, counter-intuitive and finalistic. In his account, a rational consensus reached under idealized communicative conditions *defines* validity in a truth-analogous, unconditional sense. It is important to notice that moral validity is construed as a matter of *constructing* the single right answer to a moral problem, as opposed to *finding* it. Participants in deliberation under idealized communicative conditions are given the powers of understanding and imagination required not for recognizing the single right answer but for *producing* it. This is why perfection of the procedure of intersubjective deliberation, combined with application of the principle of universalizability, is held to result in perfect moral knowledge. I see this thesis of the coincidence of human knowledge and (moral) truth as worryingly hubristic. I also think it runs counter to some of our deepest intuitions, for it eliminates the moment of receptivity to the unanticipated and to the new that we associate with the best human achievements. Furthermore, and connected with this, I regard it as finalistic, in the sense that it postulates a possible endpoint to the process of historical learning and creative human thinking, thereby denying the finitude of human knowledge, the contingency of human life and history and the creativity and freedom of human will. For, while Habermas' discourse ethics certainly allows for the thematization and critical assessment of new experiences, new needs, new interests, new ideas and so on, it rules out reflection on the *moral* validity of the procedural framework within which such matters are thematized and critically examined. In other words, it rules out moral objections to the definition of justice as the outcome of an idealized deliberative procedure in which participants agree that the norm or principle in question is equally in everyone's interests. This opens it to the accusations that it immunizes a historically and culturally specific conception of justice against attempts to rethink it in light of new experiences, events and ideas, and implies that no further learning as regards the idea of justice is necessary (Cooke 2005: 397; Cooke 2006: 129–60).

My specific objection for our present purposes is a variation of these accusations. For, just as it rules out moral criticism of the particular conception of moral validity it proposes, his language-immanent, epistemic-constructivist account of practical validity cannot accommodate the possibility that moral truth is *disclosed*, in the sense that a space opens up in which truth appears, giving rise to experiences that cause us to see ourselves, our relationships to others and our relations to the world in new, morally relevant, ways.

I have contended that truth as disclosure is the mode of truth at stake in stories of exemplarity, suggesting that translations of stories of exemplary figures and actions enable us to feel their inspirational and motivational power by opening up new spaces of the imagination in which truth appears in a new time. The space they open is one in which truth is re-presented: it appears to us anew. I have also suggested that the language-immanent—epistemic-constructivist—

conception of truth proposed by Habermas is incompatible with an understanding of truth as disclosure. If this is so, in order to develop the rich account of learning from religion towards which he gestures, and from which his theory could, in my view, profit, he will not only have to relinquish, or at least significantly modify, his thesis of ethical agnosticism, and the distinction between morality and ethics on which it is premised; he will also have to give up the language-immanent conception of truth that is the cornerstone of his account of postmetaphysical thinking.

Maeve Cooke
School of Philosophy
University College Dublin
Ireland
maeve.cooke@ucd.ie

NOTES

[1] Later in the essay, Habermas offers a fuller account of postmetaphysical thinking. Habermas 1992: 28–52.

[2] See my discussions in Cooke 2007, Cooke 2013a and in Cooke 2011.

[3] For an account of communicative rationality, see Cooke 1994.

[4] See also Cooke 1994, chapter 2.

[5] 'The Debate over the Ethical Self-understanding of the Species' in Habermas 2003b: 16–100.

[6] Translation altered. Cf. Horkheimer 1968: 263, who writes: '[Die kritische Theorie] zielt auf die Emanzipation des Menschen aus versklavenden Verhältnissen'.

[7] As we shall see Habermas defines moral validity as the outcome of an idealized procedure of deliberation in which all agree for the same reasons that a given norm or principle is equally in everyone's interests. The modern plurality of worldviews means that such rational agreement is conceivable only if participants abstract from their particular ethical conceptions, focusing instead on those interests that are universalizable. See Habermas 1993: 116–94.

[8] In Cooke 2002 I distinguish between three levels of learning: On a first level learning refers to gains in knowledge for a pre-defined purpose. Learning in this sense is essentially technical learning. On a second level it refers to beneficial changes in the self-understandings of the participants, and in their assignments of meaning and value. Learning on this level is essentially personal learning. On a third level learning refers to beneficial changes in the prevailing standards of what constitutes beneficial change on the second level. Learning on this level is essentially socio-cultural learning.

[9] See Cooke 2011: 481–2.

[10] See, for example, Cooke 2014. It is important not to confuse the question of how we come to find an argument rationally acceptable (a question of biography) with the question of what makes it rationally acceptable (a question of justification). It is also important to distinguish two positions as regards validity claims and argumentation. One holds that the rational acceptability of validity claims often depends in significant part on non-argumentative experiences. This is my position. It is what I mean when I refer to subjective experiences in specific contexts as possible co-determinants of rational accept-

ability. The other is that the experiential basis of rational acceptability renders critical evaluation of validity claims in argumentation impossible or unnecessary. This is *not* my position.

[11] I do not claim that *all* religious contents stimulate the imagination and are inspirational; equally, I do not claim that disclosure is the *only*—or even the primary—mode of religious truth.

[12] In earlier writings, I have referred to the mode of truth involved in ethical and religious experiences as 'truth as manifestation' (see, for example, Cooke 2011). The advantage of this terminology is that it highlights the way in which truth is *presented*, how it *appears* or is *shown* to us. The appearing of truth is also implied by the term 'disclosure'; this terminology has the advantage that it foregrounds the way in which the appearing of truth *opens our eyes* and brings about a *transformation* of our perceptions. Since it is this aspect that is most relevant in processes of critical evaluation, I prefer to speak of 'truth as disclosure' in the following.

[13] For the sake of simplicity, I leave aside the question of whether the terms 'critical appropriation' and 'translation' are equivalent in this context and take over Habermas' apparently interchangeable use of the terms.

[14] The following passages draw on my discussion in Cooke 2011.

[15] In the following, I pass over the specific difficulties involved in translating *stories* of exemplary figures and acts. In my view, every articulation of exemplarity is a re-presentation, in which truth contents are mediated in a particular way. In each case, the experiential component of exemplarity complicates translation. Stories are a particular mode of articulation (re-presentation). The distinctive difficulties they pose for translation are not relevant for our present purposes. It should be noted, nonetheless, that the exemplary validity of figures in literature and film is always *doubly* mediated: by the way in which it is presented in a particular literary text or film as well as by its reception by a particular reader or film-goer in a particular context (see Cooke 2013b).

[16] *Book of Genesis* 22: 1–19.

[17] This is how I interpret Søren Kierkegaard's reading of the story in Kierkegaard 2008.

[18] Ricoeur refers to truth as *manifestation* (see n. 12 above). A further important point made by Ricoeur is that the meaning of a religious work has a global shape that is engendered by the interplay between the multiple modes of articulation that produce it; thus, the meaning of the Christian New Testament emerges by way of a complex interaction between modes of articulation such as parables, proverbs, wise sayings and so on, each of which has its own distinctive features.

[19] Recall my observation that the exemplary validity of figures in literature and film is always *doubly* mediated (see n. 15 above).

[20] The argument for public evaluation in argumentation of ethical and religious claims to validity does not follow necessarily from the epistemic unreliability of ethical and religious experience, but requires a separate justification. Elsewhere I outline such a justification: see Cooke 2006: 129–33 and Cooke 2013a.

[21] I hold, however, that we can improve our knowledge of truth. For this, participation in public processes of argumentative evaluation of claims to truth is necessary, though not sufficient (Cooke 2006: 129–60).

[22] In order to show this convincingly a detailed textual analysis identifying Pamuk's use of a variety of distancing techniques would be necessary.

[23] Habermas writes: '[The] unconditional nature of moral validity claims can be accounted for in terms of the universality of a normative realm that has to be produced'. Translation altered. Cf. Habermas 1999: 301, where he speaks of *'eines herzustellenden Geltungsbereichs'*.

REFERENCES

Book of Genesis 22:1–19.
Cooke, M. (1994), *Language and Reason: A Study of Habermas's Pragmatics*. Cambridge, MA: MIT Press.
—— (2002), 'Argumentation and Transformation', *Argumentation*, 16: 79–108.
—— (2005), 'Avoiding Authoritarianism: On the Problem of Justification in Contemporary Critical Social Theory', *International Journal of Philosophical Studies*, 13: 379–404.
—— (2006), *Re-Presenting the Good Society*. Cambridge, MA: MIT Press.
—— (2007), 'A Secular State for a Postsecular Society? Postmetaphysical Political Theory and the Place of Religion', *Constellations*, 14: 224–38.
—— (2011), 'Translating Truth', *Philosophy and Social Criticism*, 37: 479–92.
—— (2013a), 'Violating Neutrality? Religious Validity Claims and Democratic Legitimacy', in C. Calhoun, E. Mendieta and J. VanAntwerpen (eds) *Habermas and Religion*. Cambridge, UK: Polity Press, 249–279.
—— (2013b), 'Representing the Good German: Philosophical Reflections', in P. O'Dochertaigh and C. Schönfeld (eds) *Representing the 'Good German' in Film and Culture*. London: Camden House, 16–28.
—— (2014), 'Truth in Narrative Fiction: Kafka, Adorno and Beyond', *Philosophy and Social Criticism* (on-line version, DOI: 0:1177/01914537145364331).
Ferrara, A. (2008), *The Force of the Example*. New York: Columbia University Press.
Gadamer, H.-G. (1979), *Truth and Method*, trans. J. Cumming and G. Barden. London: Sheed and Ward.
Habermas, J. (1984), (1987) *The Theory of Communicative Action*, vols. 1 & 2, trans. T. McCarthy. Boston: Beacon Press.
—— (1990), *Moral Consciousness and Communicative Action*, trans. C. Lenhardt and S. Weber Nicholsen. Cambridge, MA: MIT Press.
—— (1992), *Postmetaphysical Thinking*, trans. W. M. Hohengarten. Cambridge, MA: MIT Press.
—— (1993), *Justification and Application*, trans. C. Cronin. Cambridge, MA: MIT Press.
—— (1996), *Between Facts and Norms*, trans. W. Rehg. Cambridge, MA: MIT Press.
—— (1999), *Wahrheit und Rechtfertigung*. Frankfurt a.M.: Suhrkamp.
—— (2001), *The Post-National Constellation*, trans. Max Pensky. Cambridge, MA: MIT Press.
—— (2003a), *Truth and Justification*, trans. B. Fultner. Cambridge, MA: MIT Press.
—— (2003b), *The Future of Human Nature*, trans. H. Beister, M. Pensky and W. Rehg. Cambridge, MA: MIT Press.
—— (2006), *Dialectics of Secularization: On Reason and Religion*. F. Schuller and B. McNeil, San Francisco, CA: Ignatius Press.
—— (2008), *Between Naturalism and Religion*, trans. C. Cronin. Cambridge, MA: MIT Press.
—— (2010), *An Awareness of What Is Missing: Faith and Reason in a Post-Secular Age*. Cambridge, UK: Polity Press.
Horkheimer, M. (1968), *Kritische Theorie*. Frankfurt a.M.: Fischer.
—— (1982), *Critical Theory*. New York: Continuum.

Kierkegaard, S. (2008), *Fear and Trembling*. Radford, VA: A & D Publishing.
Pamuk, O. (2001), *My Name Is Red*, trans. E. Göknar. London: Faber & Faber.
Ricoeur, P. (1995), *Figuring the Sacred*, ed. M. I. Wallace, trans. D. Pellauer. Minneapolis, MN: Fortress Press.