**Schizophrenia Bulletin** vol. 38 no. 1 pp. 104–113, 2012
doi:10.1093/schbul/sbr172

# Optimization of a Goal Maintenance Task for Use in Clinical Applications

**Dori Henderson[1], Andrew B. Poppe[1], Deanna M. Barch[2,3,4], Cameron S. Carter[5,6], James M. Gold[7], John D. Ragland[5,6], Steven M. Silverstein[8,9], Milton E. Strauss[10], and Angus W. MacDonald III[1,*]**

[1]Department of Psychology, University of Minnesota, Minneapolis, MN 55455; [2]Department of Psychology, Washington University in St Louis, St. Louis, MO 63130; [3]Department of Psychiatry, Washington University in St Louis, St. Louis, MO 63130; [4]Department of Radiology, Washington University in St Louis, St. Louis, MO 63130; [5]Department of Psychiatry, University of California at Davis, Davis, CA 95616; [6]Department of Psychology, University of California at Davis, Davis, CA 95616; [7]Department of Psychiatry, University of Maryland School of Medicine, Maryland Psychiatric Research Center, Baltimore, MD 21201; [8]University of Medicine and Dentistry of New Jersey, University Behavioral HealthCare, 151 Centennial Avenue, Piscataway, NJ 08854; [9]Department of Psychiatry, University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School, Piscataway, NJ 08854; [10]Department of Psychology, Case Western Reserve University, Cleveland, OH 44106

*To whom correspondence should be addressed; Translational Research in Cognitive and Affective Mechanisms Laboratory, Departments of Psychology, University of Minnesota, N426 Elliott Hall, 75 East River Road, Minneapolis, MN 55455, US; tel: 612-624-3813, fax: 612-625-6668, e-mail and website: angus@umn.edu, www.psych.umn.edu/research/tricam/

***Background:*** **We sought to develop a Dot Pattern Expectancy task (DPX) to assess goal maintenance for use in clinical trials. Altering the standard task created 5 versions of the DPX to compare—a standard version and 4 others. Alterations in the interstimulus interval (ISI) length and the strength of a learned prepotent response distinguished the different tasks. These adjustments were designed to decrease administration time and/or improve reliability of the data.** ***Methods:*** **We determined participant eligibility in an initial session (the first of 3) using clinical interviewing tools. The initial session also included a demographic assessment and assessments of community functioning and symptom severity. All versions of the DPX were administered, across 3 sessions. Specific deficits on the context processing compared with difficulty control condition were evaluated using mixed-effects logistic regression within a hierarchical linear model.** ***Results:*** **We analyzed the data from 136 control participants and 138 participants with schizophrenia. Relative to a difficulty control condition, patients performed worse than controls on context processing conditions that required goal maintenance. ISI did not predict errors. Stronger prepotency was associated with increased errors in the difficulty control relative to context processing condition for controls, which improved the interpretability of findings for patients. Reliability was acceptable for a version of the task with a 10-minute running time.** ***Conclusions:*** **The best compromise between task duration and interpretability occurred on a version with a short ISI and a strong prepotency.**

*Key words:* goal maintenance/context processing/cognition/executive function/clinical applications/schizophrenia/translational

## Introduction

Deficits in cognition are among the most debilitating symptoms of schizophrenia.[1,2] These cognitive symptoms generally occur before the first episode of psychosis and continue throughout life.[3] To date, there are no established treatments for the cognitive symptoms associated with schizophrenia. For this reason, government and private funding agencies have prioritized research initiatives to identify the biological underpinnings of cognitive symptoms and develop therapies that target cognition. One such initiative was the Cognitive Neuroscience Treatment Research to Improve Cognition in Schizophrenia (CNTRICS) project, begun in 2007.[4] It was a direct descendent of the Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS) initiative which had focused on developing assessment tools and clinical trial designs that would be endorsed by the Food and Drug Administration (FDA) and provide pharmaceutical companies a means for gaining indications for their treatment of cognitive impairment in schizophrenia.[5] The measurement approach adopted by MATRICS involved utilizing already well-standardized clinical neuropsychological tools that were historically used in drug development trials for antipsychotics and that had known psychometric properties such as good test-retest reliability. Like MATRICS, the

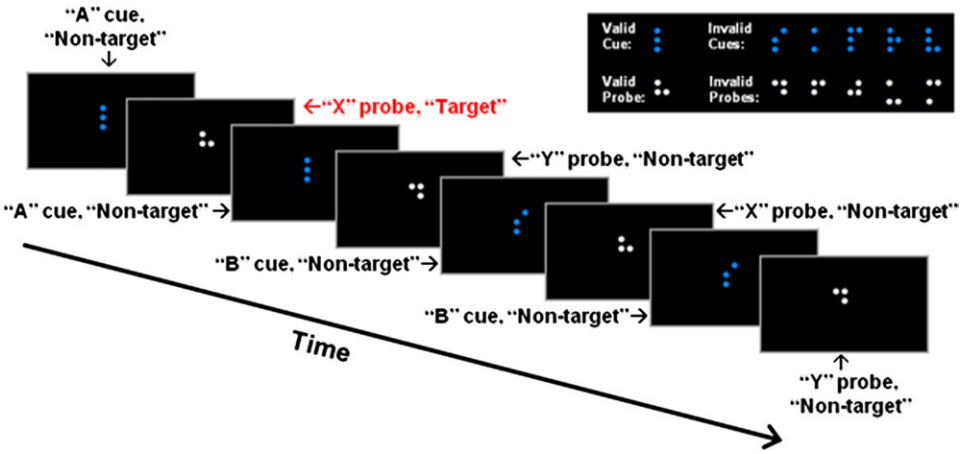


**Fig. 1.** Figure of the Dot Pattern Expectancy Task and the stimuli used. Shown is an example sequence of cue-probe stimuli and the type of response (target or non-target) a participant was required to make after each stimulus. The nomenclature for stimuli and trial types was adopted from the expectancy letter AX task. The valid cue pattern is referred to as "A" and the valid probe pattern is referred to as "X". Non-"A" cue patterns are referred to as "B"-type cues, and non-"X" probe patterns are referred to as "Y"-type probes. A target response is required to "X" when it follows "A", non-target responses are made after all other stimuli. The first pair of stimuli in the sequence represents an AX trial. The third and fourth stimuli together represent an AY type of trial, the fifth and sixth stimuli together complete a BX trial, and the seventh and eighth stimuli make up a BY type of trial. The inset shows all the valid and invalid patterns used in this study.

CNTRICS initiative followed a consensus model and involved a series of meetings and surveys to define constructs of interest and develop a measurement approach for those constructs. Unlike MATRICS, the measures under consideration for CNTRICS were from the field of cognitive neuroscience and considered by many experts to be not quite ready for "clinical trial prime time" because they were lacking formal evidence of tolerability, reliability, or correlation with clinical outcomes of interest. Additionally, most CNTRICS tasks lacked standardized administration procedures and had never been normed in large samples. CNTRICS sought to develop these cognitive neuroscience based tools because of their many potential advantages, which included the ability to evaluate discrete cognitive processes, to design tasks that could identify specific rather than generalized cognitive deficits, and the potential to link-specific cognitive deficits to neural systems in functional imaging studies and animal model systems.

The 2 major outcomes of the first set of CNTRICS meetings were (1) designation of key cognitive constructs (mechanisms) across several domains that are important in schizophrenia and (2) identification of promising measures of those constructs that could be optimized for use in clinical settings.[3,4,6,7] A third meeting focused on the measurement issues that needed to be addressed in developing these measures for use in treatment development. This special section of *Schizophrenia Bulletin* summarizes the findings of a multisite study to optimize several tasks designated by the CNTRICS initiative as ready for immediate translational development for use in clinical trials. These tasks were the Jittered Orientation Visual Integration Task (JOVI), Contrast-Contrast Effect Task, Relational and Item-Specific Encoding Task (RiSE), and the Dot Pattern Expectancy Task (DPX). This report will describe methods common to this multisite study and report on findings for the DPX task.

### *Goal Maintenance and the Dot Pattern Expectancy Task.*

A cognitive mechanism designated through the CNTRICS process as ready for immediate translational study in schizophrenia was goal maintenance.[7,8] Goal maintenance refers to the collection of cognitive processes that activate task-related goals or rules and thereby keep them represented and accessible for constraining attention, determining task-relevant information in working memory, and, ultimately, guiding behavior. One task designed to measure goal maintenance is the Dot Pattern Expectancy (DPX) task, which had strong construct validity for goal maintenance, and a task design that allowed for measurement of a specific deficit of this construct in schizophrenia and appeared to meet the other CNTRICS criteria (see figure 1, and reference[8] for a detailed discussion). The DPX is a variant of Jonathan Cohen and David Servan-Schreiber's Expectancy AX-CPT (AX-Continuous Performance Task),[9–11] a continuous performance task designed to be sensitive to context processing, ie, to a participant's ability to represent and maintain local antecedent contextual information relevant to the immediate goal. This mechanism is particularly evident in cases in which the goal state is needed to overcome an automatic, or prepotent, response.

In expectancy AX paradigms, participants view a series of cue and probe sequences, one stimulus at a time. They respond to each stimulus with either a target or nontarget response. "A" is a valid cue. "X" is a valid probe only when it follows A. All other cues (collectively referred to as "B" cues) and probes (collectively referred to as "Y" probes) are

"invalid." A "target" response is made following a "valid" probe when that probe follows a "valid" cue. That is, in letter variants of the AX task, a target response is made to X only when it follows A. Nontarget responses are appropriate to all other letters. The purpose of a "valid" and "invalid" cue (eg, A or Not A in the case of the AX task) is to manipulate the local goal state, or context, for responding to the probe (eg, when it is an X in the case of the AX task). There are 4 types of trials in the AX paradigm. "AX" trials are those in which the cue is A, and the probe is X. "AY" trials are those in which the cue is an A, but the probe letter is not X (an invalid probe). For "BX" trials, the cue is not A, but the probe is X. And finally, "BY" trials are those in which the cue is not A, and the probe is not X (ie, both probe and cue are invalid).

The DPX is identical in format to the expectancy AX task, except that the cue and probe stimuli are novel dot patterns, rather than letter stimuli. Valid and invalid cues and probes are specified, and the trial nomenclature for the DPX is adopted from the AX. Figure 1 illustrates the stimuli used, and how the trial-type nomenclature of the AX-CPT maps onto the DPX. Compared with the AX-CPT, the DPX, which uses more parametrically manageable stimuli, can require fewer trials and shorter ISIs to demonstrate a specific deficit goal maintenance (thought to be due to the way overlearned stimuli, like letters, may be stored; see Barch et al,[8] for a more complete discussion of the advantages of novel dot patterns over letter stimuli). The potentially shorter administration time for the DPX may make it more desirable than the letter AX task for clinical studies.

A key manipulation for both the letter AX and DPX paradigms is that most trials are AX trials. This establishes an expectation that X will generally follow A and encourages the development of a prepotent response bias to make target responses to letters that follow A and to all X's. The critical trials in both the expectancy AX task and DPX tasks are the AY (difficulty condition) and BX (goal maintenance condition) trial types. Individuals with intact local goal maintenance, or context processing, are expected to make more errors on AY trials relative to BX trials because good representation of the A does not reduce AY false alarms, whereas good representation of the B reduces BX false alarms. However, individuals who are less able to maintain the representation of B cues will make more errors on BX trials than those capable of maintaining this local goal state. The task design therefore permits the demonstration of a specific deficit[12,13] in goal maintenance by evaluating the participants' relative BX and AY trial performances. There is a growing body of evidence that participants with schizophrenia perform poorly on BX trials compared with controls, and their performance on AY trials is better than their performance on BX trials (for reviews, see ref.[11,14]). Thus, one way to evaluate specific deficits in patients is to compare the control and patient groups on the AY-BX difference. Performance on the expectancy AX and DPX also correlates with other tasks that require top-down control,[10] can be simulated by cognitive models that change representation and goal maintenance variables,[15,16] and is associated with activation of cognitive control regions such as the dorsolateral prefrontal cortex.[17]

Optimization of the DPX for clinical translation has yet to be accomplished. Thus, the purpose of the present study was to develop a version of the DPX task that could be used efficiently in clinical trials to assess improvement in goal maintenance. Starting with a standard version of the DPX, we created 4 systematic alterations of the standard task in order to decrease administration time or improve reliability. Each of the 5 versions were administered to psychiatrically normal participants and participants with schizophrenia. We evaluated each task on administration time, interpretability of data, ability to maintain the task's construct validity, and ability to discriminate participants with schizophrenia from controls. In the end, we determined that one version of the task most fully satisfied our requirements.

## Methods

### *Participants*

Participants for the study were recruited as part of the Cognitive Neuroscience Test Reliability and Clinical applications for Schizophrenia Consortium (CNTRaCS), which included 5 different research sites: University of California—Davis, Maryland Psychiatric Research Center at the University of Maryland, and University of Medicine and Dentistry of New Jersey, University of Minnesota—Twin Cities, and Washington University in St Louis. Participants were recruited nearly equally across the 5 different sites and were recruited from outpatient psychiatric clinics, community centers, and local settings via flyers and online advertisements. Recruiting and informed consent procedures for each site were reviewed and approved by that site's Institutional Review Boards.

Across the 5 sites, we conducted in-person screens on 141 control participants and 164 participants with schizophrenia, of which 137 control and 153 participants with schizophrenia met inclusion/exclusion criteria (see below). Six patient participants and 1 control participant were excused from the study for testing positive for drugs or alcohol. Thus, we collected behavioral data from 147 patient participants and 136 control participants. Of these individuals, 131 schizophrenia and 132 control participants completed all testing sessions, whereas 16 schizophrenia and 4 control participants completed some of the testing. After excluding for poor performance (see "Data Processing and Statistical Analyses" section below) on the DPX task, there were 138 patient participants and 136 controls participants who provided data for this article.

The inclusion and exclusion criteria were modeled after those used in the MATRICS Psychometric and Standardization Study (PASS).[18] and were designed to be similar to those that would likely be used in a study of a cognitive enhancing agent or intervention. For both

**Table 1.** Demographic and Clinical Characteristics of Participants

| | Healthy Control | | Schizophrenia Patient | | |
|---|---|---|---|---|---|
| Variable | Mean | SD | Mean | SD | Group Comparison |
| Age (in years) | 36.7 | 12.0 | 39.6 | 11.6 | $t = 1.73$, $P = .08$ |
| Gender (% males) | 55 | | 62 | | $\chi^2 = 1.17$, $P = .28$ |
| Ethnicity (% caucasian) | 54.3 | | 54.3 | | $\chi^2 = 0.14$, $P = .71$ |
| Personal education (in years) | 14.8 | 2.02 | 13.3 | 2.2 | $t = 5.89$, $P < .0001$ |
| Father education (in years) | 13.0 | 2.84 | 13.5 | 3.50 | $t = 1.35$, $P = .18$ |
| Mother education (in years) | 13.3 | 2.52 | 13.3 | 2.79 | $t = 0.09$, $P = .92$ |
| Personal SES | 38.6 | 10.3 | 26.0 | 10.2 | $t = 10.13$, $P < .0001$ |
| Parental SES | 44.4 | 12.6 | 42.8 | 15.2 | $t = 0.90$, $P = .37$ |
| BPRS average for positive symptom items | NA | | 2.19 | 1.15 | |
| BPRS average negative symptom items | NA | | 1.82 | 0.73 | |
| BPRS average for disorganized symptom items | NA | | 1.27 | 0.42 | |

*Note:* SES, socioeconomic status as measured using the using the Barratt Simplified Measure of Social Status based on the Hollingshead Index.[20] BPRS, Brief Psychiatric Rating Scale.[21–23]

controls and schizophrenia patients, the criteria included: (1) age 18–65 years, (2) no clinically significant head injury (loss of consciousness for 20 min or overnight hospitalization) or neurological disease, (3) no diagnosis of mental retardation or pervasive developmental disorder, (4) no substance dependence in the past 6 months and no substance abuse in the past month, (5) sufficient spoken English so as to be able to complete testing validity, (6) a score of 6 or higher on the Wechsler Test of Adult Reading (WTAR) as a measure of premorbid IQ,[19] (7) ability to give valid informed consent; and 8) passed alcohol and drug testing on each day of testing. Urine drug testing was conducted using the OnTrak Testcard 501 by Varian (Palo Alto, CA), which screens for cocaine, THC, methamphetamine, morphine, and amphetamine. Alcohol screenings were done using an Alcohawk Breathalyzer (<0.05%). Additional criteria for schizophrenia patients were: (1) *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*, (DSM-IV) diagnosis of schizophrenia or schizoaffective disorder (based on SCID interview, see below), (2) no medication changes in the prior month or anticipated in the upcoming month, and (3) stable outpatient or partial hospital status. Additional criteria for controls were: (1) no history of schizophrenia or any other psychotic disorder, including bipolar disorder, (2) no current major depression, and (3) no current psychotropic- or cognition-enhancing medication. The final total schizophrenia and controls groups were matched for gender, age, race, and parental socioeconomic status, which was measured using the Hollingshead Index as updated using occupational prestige ratings based on the 1989 general social survey.[20] Demographics and clinical characteristics for each group are presented in table 1.

### *Diagnosis and Clinical Assessment*

A masters level clinician conducted or supervised diagnostic assessments using the Structured Clinical Interview for DSM-IV-Text Revision[24]and the 24-item Brief Psychiatric Rating Scale.[21–23] Raters were trained by teleconferences in which ratings and anchor points for all scales were discussed, and 6 training videos were rated and discussed. Certified raters achieved agreement with the "gold" standard ratings (those of the trainers, which were highly skilled clinicians from either the St Louis or Maryland sites) for at least 6 interviews. Agreement was defined as no more than 2 items with a difference of more than 1 rating point from the gold standard. Raters added after the start of the study went through a similar process to achieve the same agreement level. To maintain reliability across the course of the study, the St Louis site created a videotaped interview to rate every 2–4 weeks, and all raters participated in a teleconference to resolve discrepancies.

### *Task and Testing Sessions*

The diagnostic interview, symptom ratings, WTAR,[19] and demographic assessment were conducted during the first session, along with 1–2 cognitive tasks. In addition, the first session included assessments of community function using the participant and informant versions of the Specific Levels of Functioning Scale[25] and a proxy measure of function, the Brief University of California, San Diego, Performance-based Skills Assessment (UPSA-B).[26–28] Participants then completed between 2 and 3 additional cognitive testing sessions within approximately 1 month. Across these sessions, participants performed 5 versions of the DPX, 1 version each of the JOVI

**Table 2.** Task Parameters, Performance, and Reliability on 5 Versions of the DPX

| | | | Schizophrenia Patients | | Controls | |
|---|---|---|---|---|---|---|
| Version | Trial-Type | Number of Trials (%) | % Errors (SD) | α | % Errors (SD) | α |
| Long forms (ISI 4000 ms) | | | | | | |
| Form 1 | AX | 88 (68.75) | 9.1 (8.1) | .87 | 3.7 (4.2) | .78 |
| | AY | 16 (12.5) | 20.1 (17.0) | .71 | 10.3 (10.2) | .47 |
| | BX | 16 (12.5) | 19.9 (21.2) | .83 | 8.3 (12.5) | .74 |
| | BY | 8 (6.25) | 4.2 (8.5) | .32 | 0.9 (3.7) | .15 |
| Form 2 | AX | 80 (62.5) | 10.2 (10.2) | .90 | 4.3 (5.1) | .80 |
| | AY | 24 (18.75) | 14.1 (14.6) | .80 | 7.7 (9.2) | .68 |
| | BX | 16 (12.5) | 17.3 (19.7) | .82 | 9.5 (13.8) | .76 |
| | BY | 8 (6.25) | 5.9 (9.6) | .27 | 2.1 (6.5) | .43 |
| Form 3 | AX | 76 (59.38) | 10.5 (10.1) | .90 | 5.2 (6.9) | .88 |
| | AY | 24 (18.75) | 14.3 (14.2) | .78 | 6.6 (8.1) | .64 |
| | BX | 20 (15.63) | 20.5 (22.0) | .85 | 9.0 (12.9) | .80 |
| | BY | 8 (6.25) | 5.4 (10.3) | .43 | 1.8 (5.0) | .14 |
| Short forms (ISI 2000 ms) | | | | | | |
| Form 1 | AX | 88 (68.75) | 9.1 (9.2) | .90 | 2.8 (3.9) | .80 |
| | AY | 16 (12.5) | 18.1 (15.1) | .65 | 9.8 (5.0) | .39 |
| | BX | 16 (12.5) | 20.4 (20.4) | .79 | 7.8 (4.8) | .53 |
| | BY | 8 (6.25) | 7.1 (11.1) | .28 | 1.8 (7.4) | .21 |
| Form 2 | AX | 80 (62.5) | 9.6 (9.9) | .90 | 3.2 (12.3) | .84 |
| | AY | 24 (18.75) | 14.2 (13.1) | .74 | 6.2 (6.4) | .58 |
| | BX | 16 (12.5) | 20.2 (19.8) | .80 | 8.2 (5.0) | .73 |
| | BY | 8 (6.25) | 7.1 (10.3) | .27 | 2.1 (4.8) | .38 |

*Note:* ISI, interstimulus interval (time between the offset of the cue and the onset of the probe, in milliseconds); Reliability is measured by Cronbach's α for available participants; For all versions of the task, the number of trials was 128, cue length was 1000 ms, probe length was 500 ms, and intertrial interval was 1200 ms. Long form 1 represented the standard version of the task. The other task versions were developed by altering the standard version's ISI and critical trial (AY or BX) frequencies. Long forms 2 and 3 were created by increasing the frequency of one (long form 2; AY trials) or both (long form 3; AY and BX trials) critical trials. Short form 1 was created by reducing the ISI, whereas short form 2 reduced ISI and increased AY trial frequency. AY or BX trial frequency increases were accommodated by a reduction in AX trial frequency (ie, the expectancy manipulation).

and Spatial Offset Visual Integration (SOVI) task, 2 versions of the Contrast Contrast Effect (CCE) task, and 3 versions of the RiSE task. Tasks other than DPX are described elsewhere in this special section. Within a task type (eg, DPX, CCE, RISE), version was counterbalanced using a latin-squares design. Participants never did more than one version of the RISE or CCE in a single session and never more than 2 versions of the DPX in a single session.

The goal of our project was to develop a version of the DPX that would require minimal administration time but maintain good construct validity and interpretability. As specified in table 1, we created and tested 5 versions of the DPX that varied in the duration of their cue-probe interstimulus interval (ISI) or in the number of critical trials (AY or BX). Previous research has suggested that that increasing ISIs from 1000 to 5000 ms results in a task that is more sensitive to goal maintenance deficits in people with schizophrenia. However, reducing ISI could potentially reduce administration time. The purpose of this manipulation in the current study was to find an ISI that would maintain sensitivity while minimizing task administration time. Increasing the number of critical trials would tend to reduce floor and ceiling effects and increase the reliability—particularly of the AY condition (though achieving this advantage would be at the expense of AX trials, which could reduce the difficulty of the AY condition).

The 5 versions of the DPX consisted of a standard "long" version (LF1) and 4 others (2 additional "long" versions and 2 "short" versions). The versions were created by altering the standard ISI length (4000 for "long" versions and 2000 ms for "short" versions) and increasing the number of AY or BX (with a concomitant decrease in AX trial frequency). We hoped to arrive at a version that maximized critical trial reliability, while maintaining at least a relatively strong expectancy manipulation (which is presumed to maintain the strength of learned prepotent responding to A that underlies AY error rates generation, with more AX trials associated with a greater response prepotency). Essential features of the 5 DPX tasks are summarized in table 2.

### *Data Processing and Statistical Analyses*

Two hundred and eighty-three qualifying participants (146 patients and 137 controls) were tested, of whom 9 (8 patients and 1 control) were removed for poor

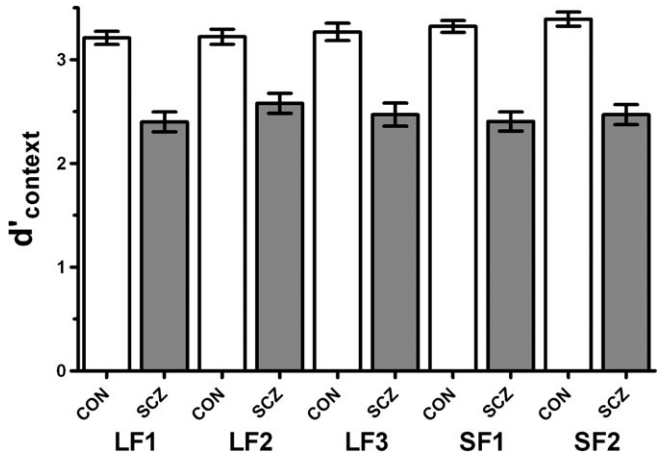


**Fig. 2.** Mean and SE of $d'_{context}$ for patient (SCZ) and control (CON) groups for each of 5 DPX tasks. LF1 (long form 1) and SF1 (short form 1) had 69%, LF2 (long form 2) and SF2 (short form 2) had 62%, and LF3 (long form 3) had 58% AX trials.

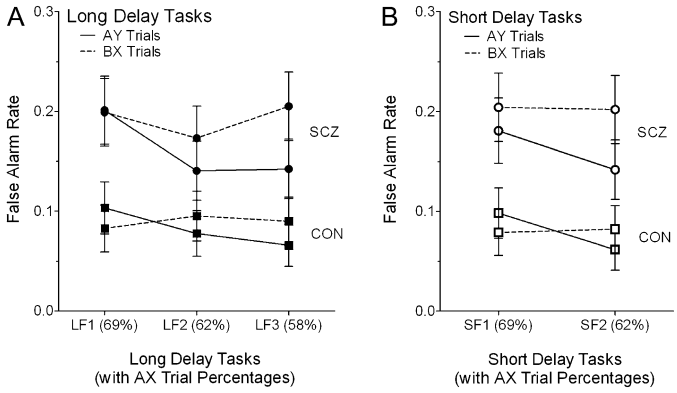


**Fig. 3.** Average false alarm rates and SEs on AY and BX trials as a function of interstimulus interval. Shown are average false alarm rates for AY (solid lines) and BX (dashed lines) trial types for the long delay (panel A) and short delay (panel B) tasks. Circles represent performance for patient participants and squares represent the performance of controls.

performance. Poor performance was defined as error rates greater than or equal to 56% on AX trials (indicating worse than chance performance), 100% on AY or BX trials, or 50% on BY trials (low BY trial accuracy indicates a participant failed to grasp the task). We calculated $d'_{context}$ for remaining participants in each of the 5 DPX conditions. $D'_{context}$ is calculated using AX hits and only BX false alarm rates rather than all false alarms to emphasize the condition that requires ongoing maintenance of the A context to perform accurately.[9] A small constant was used to estimate $d'_{context}$ in the case of 100% accuracy on either trial-type.[29] A repeated measures ANOVA with a Huynh-Feldt correction for violations of sphericity compared performance across the 5 tasks for subjects with complete data.

Although $d'_{context}$ provides an estimate of sensitivity corrected for response bias, it does not distinguish specific and generalized performance deficits in patients.[11–13] Instead, a specific deficit in context processing can be examined through an analysis of differential response patterns across BX and AY trials for patients and controls (with patients performing worse on BX trials compared with AY). Therefore, we also used a mixed-effects logistic regression on trial-to-trial accuracy data for all tasks within a linear model (HLM)[30] approach implemented in R[31]. This approach, which uses a $z$ test as the inferential statistic to test differences in the estimated threshold between accurate and inaccurate performance, allowed us to model trial-by-trial binary outcomes (accurate vs inaccurate) without relying on parametric assumptions. The approach also allowed us to contrast performance on the AY condition (difficulty comparison) and the BX context processing (goal maintenance) conditions across patient and control groups. That is to say, the effects of group, trial-type, and version were included in the logistic regression models when they improved the fit of the model as measured by a "decrease" in the Akaike's or Bayesian Information Criterion (AIC and BIC). These criteria both credit the statistical goodness of fit of differing models in the same way ($-2\mathrm{L}m$, where L$m$ is the maximized log-likelihood of the model). They differ in that AIC penalizes for each parameter used to achieve this fit at $2m$, where $m$ is the number of parameters used to achieve fit. BIC penalizes more steeply for each parameter, at $\ln(n)m$ where $n$ is the sample size and is therefore a more conservative criterion.

## Results

### *Version Selection*

Figure 2 shows the means and SEs for $d'_{context}$ for patient and control participants for each of the 5 tasks. A repeated-measures ANOVA using subjects who had valid data for all 5 versions showed that patient participants performed worse than controls for all 5 versions ($F_{1,196} = 37.6$, $P = 1.6 \times 10^{-4}$). We observed no significant interaction between group and version ($F_{3.8,755} = 2.03$, $P = .09$). The effect sizes ranged from .67 for LF2 condition to 1.07 for LF1 condition. The results demonstrate that $d'_{context}$ is not particularly sensitive to differences between the 5 tasks.

$D'_{context}$ is limited in that it does not distinguish between goal maintenance deficits and general impairments. To evaluate whether these deficits were indicative of a "specific" failure of goal maintenance, we considered the relative deficit of patients on BX compared with AY trials. As presented in table 2 and figure 3A and B, across all 5 versions, AY trials were harder than AX and BY trials ($z = 20.18$ and $z = 14.49$, both $P < 2.0 \times 10^{-16}$) and were generally slightly easier than BX trials ($z = -2.00$, $P = .046$). Overall, schizophrenia patients performed worse than controls across all trial types. There was marked evidence of a deficit on the AY difficulty control condition

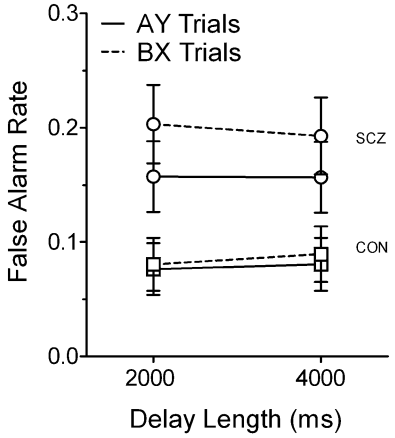


**Fig. 4.** Effect of the interstimulus interval length on AY and BX trial performance for the patient and control groups on task SF1. The figure shows average false alarm rates and SEs for AY (solid lines) or BX (dashed lines) trial types as a function of short (2000 ms) or long (4000 ms) ISIs. Circles indicate performance for patient participants and squares represent the performance of controls.

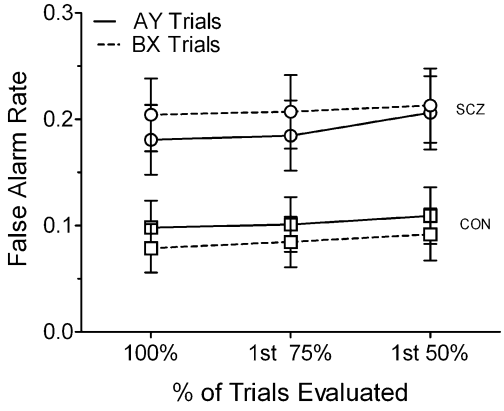


**Fig. 5.** Average false alarm rates and SEs on AY and BX trials calculated for 100%, first 75%, or first 50% of trials for SF1. Shown are average false alarm rates averaged across AY (solid lines) and BX (dashed lines) for the patient (circles) and control (squares) groups.

(controls vs patients $z = -7.61$, $P = 2.8 \times 10^{-14}$). However, performance on AY trials was relatively spared as patients performed comparatively worse than controls on BX trials ($z = -3.48$, $P = .0005$) as well as AX ($z = -4.41$, $P = 1.1 \times 10^{-5}$) and BY ($z = -3.81$, $P = 1.4 \times 10^{-4}$) trials.

The reliability (internal consistency) for all trial types on all versions is also presented on table 2, indicating generally better, and acceptable, levels of internal consistency for patients on AX, AY, and BX trials, with low internal consistency for rarer BY trials, the manipulation check. For readers interested in examining item reliability, which is independent of number of trials, we include in table 1S of the online supplementary materials, the average interitem correlations ($r_{ij}$) for the different trial-types for each of the 5 tasks.

We next tested the effect of the ISI, which improved model fit according to AIC ($\Delta$AIC [$\Delta df = 8$] = $-85$, $P < 2.2 \times 10^{-16}$) and BIC ($\Delta$BIC [$\Delta df = 8$] = $-5$). However, the improved fit was due to a small accuracy increase on AX trials at the shorter ISI (see table 2). There were actually no delay-related effects on AY, BX, or BY trials and no interactions between group and delay on these trial types (figures 3 and 4). For this reason, we focused on the more time efficient short-delay versions of the DPX (SF1 and SF2) for subsequent analyses.

The next model considered the effect of altering the proportion of AX trials (69% vs 63%—short form 1 and short form 2, respectively). Including this factor in the model improved fit according to AIC ($\Delta$AIC [$\Delta df = 8$] = $-31$, $P < 1.02 \times 10^{-7}$) but not according to the more conservative BIC ($\Delta$BIC [$\Delta df = 8$] = 41), suggesting that the proportion of AX trials was likely, but not certain, to play a role in performance. In this case, for controls, the 69% condition was harder for AY trials ($z = 5.023$, $P = 5.09 \times 10^{-7}$), whereas the AX, BX, and BY conditions were relatively but not absolutely easier ($z$'s $< -2.2$, $P$'s $< .03$), which was a desirable property because it suggests that other things being equal the AY condition was further from the ceiling and therefore potentially more sensitive to group differences. There was no interaction between group and proportion for any trial type ($z$'s $< 1.6$, $P$'s $> .10$).

Thus, there were 2 advantages favoring the short ISI/69% AX condition (SF1) as the final candidate task: (1) the running time of SF1 was 33% shorter than for any of the longer versions without consequence to the AY vs BX interaction with group and (2) for controls, the AY condition was harder than the BX condition ($z = 2.3$, $P = .02$) thereby increasing the likelihood that this AY comparison condition could be used to control for generalized deficits. In fact, this turned out to be the case as the group by trial type (AY vs BX) interaction was both significant ($z = -3.07$, $P = .002$) and interpretable as a fan-shaped interaction (the AY condition was the more difficult for controls, whereas the group difference was biggest on BX trials; see short form 1 in table 2).

We used a repeated measures ANOVA to evaluate reaction time differences between the patients and controls on accurate AY and BX trials for SF1 to determine whether accuracy differences may be the result of group differences in a speed-accuracy trade-off. Our analyses indicated a main effect of group ($F_{1,260} = 54.26$, $P < .001$), with controls responding faster than patients. The mean response latency (±SD) was 438 ms (81 ms) for controls and 518 ms (119). Both groups responded faster on BX trials compared with AY trials (controls: AY = 575 [83] ms, BX = 359 [82] ms; patients': AY = 647 [96] ms, BX 441 [137] ms). We found no interaction between group and trial type ($F_{1,260} = .547$, $P = .46$), suggesting that speed-accuracy trade-off differences did not account for the group differences in accuracy described above.

We evaluated whether eliminating trials could further shorten the SF1 version. To assess this, we looked at BX

and AY trial performance for different proportions of trials. Figure 5 shows average false alarm rates on AY or BX trials calculated for all trials, the first 75%, and the first 50% of trials of the SF1 task. Power to detect the group by trial-type (AY vs BX) interaction declined as data was removed, and so the effect sizes were smaller: with 100% of trials $z = -3.07$ ($P = .002$), with 75% $z = -2.3$ ($P = .02$), and with 50% $z = -1.4$, ($P = 0.17$). Although these models could not be directly compared because they were not nested, it appeared that the precision with which ability was measured continued to increase across trials, whereas the nature of the interaction did not change across the duration of the task. Since there was no evidence of an asymptote, the current data do not rule out the possibility of a larger effect if the number of trials were further increased.

An interesting aspect of the design was that participants completed trials in a fixed order that was pseudorandomly generated to help optimize prepotency effects. As a result, they received a similar sequence of consecutive AX trials before receiving an AY, BX, or BY trial. Consistent with the goals of this ordering, the number of preceding AX trials appeared to be relevant to task performance, in that the model with this information improved fit of SF1 data according to AIC ($\Delta$AIC [$\Delta df = 12$] $= -7$, $P = .002$) though not according to BIC ($\Delta$BIC [$\Delta df = 12$] $= 78$). This appeared to be due to somewhat better BX relative to AY performance after 3 AX repetitions ($z = 2.88$, $P = .004$), largely, but not significantly more among controls. These findings suggest that an expectancy manipulation that generates the prepotency to respond with a target response following an A can operate locally—across just a few trials. The potential influence of local prepotency effects—runs of alternating responses to AX trials—is an interesting future direction for task development.

## Discussion

The aim of this study was to develop a goal maintenance task that was optimized for use in clinical trials. The impetus for this work was the CNTRICS initiative, which had previously identified a set cognitive constructs and measures that were ready for translation from experimental to clinical applications and if further honed might be used to test new treatments.[3,4,6–8] Goal maintenance as measured by the DPX was among those identified by CNTRICS and was the focus of this study. To this end, we demonstrated a variant of the DPX task that had a shorter administration time could still convincingly measure specific deficits in goal maintenance.

We evaluated schizophrenia patients and controls across multiple sites on 5 variants of the DPX task to determine which had the most optimal psychometric properties for clinical trials. We found the best compromise between task duration and conceptually grounded interpretability on a variant with a short ISI and a strong prepotency. The effect size increased over the course of all trials, indicating that the full 10-minute version of the task was the most sensitive to group differences. The internal consistency of the AX, AY, and BX conditions on this task version were at acceptable levels for patients, though not for controls.[3] Other internal consistency estimates, disattenuated for the number of trials, suggested AY and BX reliabilities were generally in the low range of acceptable scores.

Two recent studies evaluated the psychometric characteristics of the expectancy AX and the DPX tasks. For the AX task, coefficient alpha for error rates were .80 for AX, .71 for AY, and .90 for BX trials across a mixed group of 63 schizophrenia patients, their siblings, and controls.[8] For the DPX task, alphas were .94 for AX, .65 for AY, .87 for BX, and .55 for BY across 95 schizophrenia patients and controls with a similar number of trials and a 4 second cue-probe ISI.[14] The internal consistency was similar for patients when assessed alone. Thus, the current results are the third study demonstrating that internal consistency of the key AX and BX conditions in this kind of task can be comparable to those of clinical neuropsychological tasks, at least among patients with schizophrenia. The AY condition was found to be somewhat less internally consistent, but at .65, this was (and has consistently been) above the minimum recommended by expert consensus.[3] The BY condition had reduced reliability, in large part due to the ceiling and near-ceiling performance of many participants on this condition (note that the BY trial type is the easiest and was intended to serve as a check that participants understand the task). The low reliability for this trial type should not, therefore, detract from the task's utility.

It is possible that increasing the AX prepotency even beyond 69% (the highest of the DPX versions tested) would further increase strength of the AY vs BX interaction. Increased interaction strength might be desirable to detect more subtle effects, for example, changes in performance as a function of treatment. However, this would require more testing time, which imposes an opportunity cost. A second limitation is that although we tested 2 ISIs, it is unknown whether shortening the interval further could achieve still greater efficiencies without losing the underlying effect. Third, it is useful to note that trials were in a fixed order to maximize sensitivity to individual differences while avoiding a priori ordering effects. One might imagine, given the examination of local AX prepotency effects, that a stimulus order that further optimized these prepotency effects might further increase the efficiencies of the task. Finally, our findings replicate previous studies showing that patients sometimes do worse on AY trials relative to controls, even though the magnitude of this group difference was less than for BX trials. The AY trials are supposed to be a difficulty control condition. This is a limitation as it decreased the potential strength of the interaction between group and trial type. It is important to recall that the AY vs BX comparison does not "get rid" of patients' impairment (although that was an early goal[9,16]) but may

be more useful as a means to discern whether there is a context processing deficit over-and-above this deficit. Thus, a BX condition that is easier for controls than the AY condition helps provide this interpretive leverage, but it does not eliminate patients' overall difficulties on the task.

CNTRICS identified a set of cognitive neuroscience–based tools to be used in the measurement of specific cognitive processes that may be impaired in schizophrenia for the purpose of targeting those processes in clinical trials. The form of goal maintenance known as context processing has been shown to depend critically on prefrontal cortical functioning, and failures of functioning in this brain region correspond to an increase in context processing deficits and disorganization symptoms.[30] The current evaluation suggests that a version of the DPX—one that takes 10 minutes of running time to conduct and appears to have adequate internal consistency, performance off of ceiling, and an interpretable pattern of errors in patients with schizophrenia—may be useful for efficiently measuring context processing. While test-retest reliability and practice effects have not been determined with this version, this will be an important question to examine in subsequent studies.

## Funding

National Institute of Mental Health (5R01MH084840-03 to D.M.B., 5R01MH084826-03 to C.S.C., 5R01MH084828-03 to S.M.S., 5R01MH084821-03 to J.M.G., 5R01MH084861-03 to A.W.M.).

## Supplementary Material

Supplementary material is available at http://schizophreniabulletin.oxfordjournals.org.

## Acknowledgments

We acknowledge the hard work and dedication of staff at each of the CNTRaCS sites. We also thank our participant for their essential contributions to this study. Financial Disclosures: D.H. and A.B.P. have no currently active grant or contract support from private or public sources. D.M.B. has received grants from the National Institute of Health (NIH), National Alliance for Research in Schizophrenia and Depression (NARSAD), Allon, Novartis, and the McDonnell Center for Systems Neuroscience. C.S.C. has received research grants from the National Institute of Mental Health, National Institute on Drug Addiction (NIDA), the Robert Wood Johnson Foundation and from Glaxo Smith Kline and has been an external consultant for Roche, Servier, Lilly, Merck, and Pfizer. J.M.G. has received grants from NIH, receives royalty payments from the BACS, and has consulted with Pfizer, Merck, Astra Zenaca, Solvay, and Glaxo Smith Kline. J.D.R. has received research grants from the NIH, NARSAD and the Robert Wood Johnson Foundation. S.M.S. has received research grants from NIH, NARSAD, Pfizer, Eli Lilly, and AstraZeneca. M.E.S. has no currently active grant or contract support from private or public sources. A.W.M. has received research grants from the NIH and NARSAD. The authors have declared that there are no conflicts of interest in relation to the subject of this study.

## References

1. Heinrichs RW. The primacy of cognition in schizophrenia. *Am Psychol*. 2005;60:229–242.
2. Green MF. *Schizophrenia From A Neurocognitive Perspective: Probing The Impenetrable Darkness*. Boston, MA: Allyn and Bacon; 1998.
3. Barch DM, Carter CS. Committee tCE. Measurement issues in the use of cognitive neuroscience tasks in drug development for impaired cognition in schizophrenia: a report of the second consensus building conference of the CNTRICS initiative. *Schizophr Bull*. 2008;34:613–618.
4. Carter CS, Barch DM. Cognitive neuroscience-based approaches to measuring and improving treatment effects on cognition in schizophrenia: the CNTRICS initiative. *Schizophr Bull*. 2007;33:1131–1137.
5. Green MF, Nuechterlein KH, Gold JM, et al. Approaching a consensus cognitive battery for clinical trials in schizophrenia: the NIMH-MATRICS conference to select cognitive domains and test criteria. *Biol Psychiatry*. 2004;56:301–307.
6. Barch DM, Carter CS, Arnsten A, et al. Selecting paradigms from cognitive neuroscience for translation into use in clinical trials: proceedings of the third CNTRICS meeting. *Schizophr Bull*. 2009;35:109–114.
7. Carter CS, Barch DM, Buchanan RW, et al. Identifying cognitive mechanisms targeted for treatment development in schizophrenia: an overview of the first meeting of the cognitive neuroscience treatment research to improve cognition in schizophrenia initiative. *Biol Psychiatry*. 2008;64:4–10.
8. Barch DM, Berman MG, Engle R, et al. CNTRICS final task selection: working memory. *Schizophr Bull*. 2009;35:136–152.
9. Servan-Schreiber D, Cohen J, Steingard S. Schizophrenic deficits in the processing of context: a test of a theoretical model. *Arch Gen Psychiatry*. 1996;53:1105–1112.
10. Cohen JD, Barch DM, Carter C, Servan-Schreiber D. Context-processing deficits in schizophrenia: converging evidence from three theoretically motivated cognitive tasks. *J Abnorm Psychol*. 1999;108:120–133.
11. MacDonald AW III. Building a clinically relevant cognitive task: case study of the AX paradigm [published online ahead of print May 16, 2008]. *Schizophr Bull*. 2008;34:619–628.
12. Strauss ME. Demonstrating specific cognitive deficits: a psychometric perspective. *J Abnorm Psychol*. 2001;110:6–14.
13. Chapman LJ, Chapman JP. Problems in the measurement of cognitive deficit. *Psychol Bull*. 1973;79:380–385.
14. Jones JAH, Sponheim SR, MacDonald AW, III. The dot pattern expectancy (DPX) task: reliability and replication of deficits in schizophrenia. *Psychol Assess*. 2010;22:131–141.
15. Cohen JD, Servan-Schreiber D. Context, cortex and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psycholl Rev*. 1992;99:45–77.

16. Braver T, Cohen J. Working memory, cognitive control, and the prefrontal cortex: computational and empirical studies. *Cogn Process*. 2001;2:25–55.
17. Minzenberg MJ, Laird AR, Thelen S, Carter CS, Glahn DC. Meta-analysis of 41 functional neuroimaging studies of executive function in schizophrenia. *Arch Gen Psychiatry*. 2009;66:811–822.
18. Nuechterlein KH, Green MF, Kern RS, et al. The MATRICS consensus cognitive battery: part 1: test selection, reliability, and validity [published online ahead of print January 02, 2008]. *Am J Psychiatry*. 2008;165:203–213.
19. Wechsler D. *Wechsler Test Of Adult Reading*. San Antonia, TX: The Psychological Corporation; 2001.
20. Hollingshead AD, Redlich FC. *Social Class And Mental Illness.* New York, NY: Wiley; 1958.
21. Ventura J, Green MF, Shaner A, Liberman RP. Training and quality assurance on the Brief Psychiatric Rating Scale: the "drift busters". *Int J Methods Psychiatry Res*. 1993;3:221–226.
22. Ventura J, Lukoff D, Nuechterlein KH, Liberman RP, Green MF, Shaner A. Brief Psychiatric Rating Scale (BPRS) exanded version: scales, anchor points, and administration manual. *Int J Psychiatr Methods*. 1993;3:227–243.
23. Overall JE, Gorham DR. The brief psychiatric rating scale. *Psychol Rep*. 1962;10:799.
24. First MB, Spitzer RL, Miriam G, Williams JBW. *Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Non-patient Edition (SCID-I/NP)*. New York, NY: Biometrics Research, New York State Psychiatric Institute; 2002.
25. Schneider LC, Struening EL. SLOF: a behavioral rating scale for assessing the mentally ill. *Soc Work Res Abstr*. 1983;19:9–21.
26. Patterson TL, Goldman S, McKibbin CL, Hughs T, Jeste DV. UCSD performance-based skills assessment: development of a new measure of everyday functioning for severely mentally ill adults. *Schizophr Bull*. 2001;27:235–245.
27. Twamley EW, Doshi RR, Nayak GV, et al. Generalized cognitive impairments, ability to perform everyday tasks, and level of independence in community living situations of older patients with psychosis. *Am J Psychiatry*. 2002;159:2013–2020.
28. Harvey PD, Velligan DI, Bellack AS. Performance-based measures of functional skills: usefulness in clinical treatment studies. *Schizophr Bull*. 2007;33:1138–1148.
29. Nuechterlein KH. Vigilance in schiophrenia and related disorders. In: Steinhauer SR, Gruzelier JH, Zubin J, eds. *Handbook of Schiophrenia, Volume 5: Neuropsychology, Psychophysiology and Information Processing*. New York, NY: Elsevier; 1991.
30. MacDonald AW, III, Carter CS, Kerns JG, et al. Specificity of prefrontal dysfunction and context processing deficts to schizophrenia in a never medicated first-episode psychotic sample. *Am J Psychiatry*. 2005;162:475–484.
31. R Development Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2011. ISBN 3-900051-07-0, URL http://www.R-project.org/.