American Journal on Intellectual and Developmental Disabilities

Direct Measures of Medication Effects: Exploring the Scientific Utility of Behavior-Analytic Assessments
--Manuscript Draft--

Manuscript Number:	AJIDD-D-20-00027R2
Article Type:	Research Report
Keywords:	generalizability; reliability; progressive ratio; demand assessment; psychotropic medication
Corresponding Author:	Blair Lloyd Vanderbilt University Nashville, TN UNITED STATES
First Author:	Blair Lloyd
Order of Authors:	Blair Lloyd
	Emily Weaver
	Jessica Torelli
	Marney Pollack
	Sunya Fareed
	Angela Maxwell-Horn
Manuscript Region of Origin:	UNITED STATES
Abstract:	The purpose of the current study was to explore the scientific utility of two behavior analytic assessments (i.e., progressive ratio and demand assessments) for psychotropic medication evaluation. For a sample of 23 children with disabilities who were prescribed medication, we conducted a series of generalizability and optimization studies to identify sources of score variance and conditions in which stable estimates of behavior can be obtained. To inform construct validity, we calculated correlations between scores from each assessment and those from a standardized behavior rating scale (ABC-2). Results offer initial support for the scientific utility of progressive ratio scores. More research is needed to evaluate sensitivity to change and construct validity of scores from these and other behavior analytic assessments.

Abstract

The purpose of the current study was to explore the scientific utility of two behavior analytic assessments (i.e., progressive ratio and demand assessments) for psychotropic medication evaluation. For a sample of 23 children with disabilities who were prescribed medication, we conducted a series of generalizability and optimization studies to identify sources of score variance and conditions in which stable estimates of behavior can be obtained. To inform construct validity, we calculated correlations between scores from each assessment and those from a standardized behavior rating scale (ABC-2). Results offer initial support for the scientific utility of progressive ratio scores. More research is needed to evaluate sensitivity to change and construct validity of scores from these and other behavior analytic assessments.

Keywords: generalizability, reliability, progressive ratio, demand assessment, psychotropic medication

Direct Measures of Medication Effects:

Exploring the Scientific Utility of Behavior Analytic Assessments

The prevalence of psychotropic medications among school-age children has increased markedly over the past 20 years (Carlson, 2019; Olfson et al., 2015), particularly among children with disabilities. As many as 50%–70% of children with autism spectrum disorders, attention deficit/hyperactivity disorder (ADHD), and other emotional/behavioral disorders take one or more psychotropic medication (Angold et al., 2000; Mandell et al., 2008; Mattison et al., 2014; Ryan et al., 2008; Spencer et al., 2013), many of which are prescribed to address behavioral concerns. Coupled with rising rates of off-label prescribing and polypharmacy (McLaren et al., 2018; Vitiello, 2017; Zito et al., 2008), these trends magnify the need to understand whether and how these medications produce therapeutic effects.

To evaluate whether medication-based treatments are effective for children with disabilities, prescribing clinicians commonly rely on informal reports by caregivers or global impressions of symptom improvement (e.g., Clinical Global Impressions [CGI] scale [Guy, 1976]; Vitiello, 2017). Researchers commonly rely on third party reports of child behavior by parents or other caregivers (Aman et al., 2004; Volpe et al., 2005; Zarcone et al., 2008). The Aberrant Behavior Checklist (ABC; Aman et al., 1985), for example, is a standardized behavior rating scale used to monitor behavioral or psychiatric problems for individuals with disabilities. The ABC is highly regarded by experts (Rush & Frances, 2000) and has been used as a primary outcome measure in numerous clinical trials, some of which led to approvals of psychotropic medications by the Food and Drug Administration for children with autism (e.g., Marcus et al., 2009; McCracken et al., 2002). Practical advantages of indirect assessments notwithstanding, these methods provide information on other people's perceptions of the extent to which

medication has impacted child behavior.

In contrast to third party reports, behavior analytic assessments are designed to identify

and describe behavior-environment interactions (i.e., behavioral processes) by directly observing

child behavior as environmental variables are systematically programmed. Behavior analytic

researchers have hypothesized that psychotropic medication effects can be understood in terms

of basic behavioral processes (e.g., Cox & Virues-Ortega, 2016; Lloyd et al., 2016; Thompson et

al., 2007; Valdovinos & Kennedy, 2004). Specifically, these researchers have posited that

psychotropic medications may act as motivating operations, such that they temporarily increase

or decrease the value of a reinforcer, and thus the likelihood of behaviors that produced that

reinforcer in the past (Laraway et al., 2003). Results of several empirical studies have offered

preliminary support for this conceptualization. For example, there is evidence suggesting

stimulant medication changes the value of positive reinforcers (e.g., edibles, social attention;

Northup et al., 1997; Dicesare et al., 2005) and that atypical antipsychotic medication decreases

the value of negative reinforcers (e.g., escape from non-preferred tasks; Crosland et al., 2003;

Danov et al., 2012; Zarcone et al., 2004).

Assessments that inform behavioral processes could be useful for determining child

response to intervention—including medication—as objective measures that supplement or

replace third-party report (Northup & Gulley, 2001; Volpe et al., 2005; Weeden et al., 2009). But

identifying relevant behavioral processes impacted by medication has further important clinical

implications. For one, it might help identify behavioral indicators that predict responsiveness to

medication, potentially explaining why two children with the same diagnostic profile can

respond differently to the same treatment (Kollins et al., 2000). For another, identifying

underlying behavioral processes by which medications produce a therapeutic effect could inform

ways to better integrate medication-based treatment with other interventions focused on skill

building (Lloyd et al., 2016; Volpe et al., 2019). This implication is especially important in light

of recommendations supporting psychopharmacologic intervention as an addition to—as

opposed to a replacement for—behavioral or other psychosocial interventions (AACAP, 2012;

Rush & Frances, 2000).

Two behavioral assessments that inform positive and negative reinforcer value,

respectively, are progressive ratio and demand assessments. Progressive ratio schedules are those

in which reinforcers are delivered contingent on a fixed number of responses that systematically

increases within a session (Roane, 2008). Measures of responding are used to determine the

amount of effort a person is willing to put forth to earn a given reinforcer. One common index of

responding is known as the breakpoint: the response value of the last schedule completed in a

session. Applied researchers have used progressive ratio assessments to compare positive

reinforcer value among multiple stimuli for a single participant (e.g., Roane et al., 2001). This

information is then used to identify the most potent positive reinforcers to incorporate in

behavioral interventions. Though less common, progressive ratio schedules can also be repeated

with the same stimuli to evaluate change in positive reinforcer value over time or between

conditions (e.g., Chelonis et al., 2011).

Demand assessments can be used to inform the value of negative reinforcement (i.e.,

escaping or avoiding unpleasant conditions). Similar to escape conditions of a functional analysis

(Thomason-Sassi et al., 2011), sessions involve continuous presentation of non-preferred task

demands to determine how long a person will tolerate these demands without engaging in

problem behavior. Latencies to problem behavior are used as an index of task aversiveness, with

shorter latencies indicating lower demand tolerance (i.e., increased negative reinforcer value;

Call et al., 2009; 2016). Similar to progressive ratio assessments, demand assessments have been used to compare negative reinforcer value among multiple task demands for a single participant (Call et al., 2009; 2016), but these and similar assessments can also be repeated with the same task demand to evaluate changes in negative reinforcer value over time or between conditions (e.g., Crosland et al., 2003; Zarcone et al., 2004). Both progressive ratio and demand assessments have been recommended by behavior analytic researchers for purposes of medication evaluation and have potential to inform changes in reinforcer value between medication conditions (Carlson et al., 2012; Crosland et al., 2003; Roane, 2008; Zarcone et al., 2004). However, few studies have applied these assessments for this purpose, and to our knowledge, none have evaluated the reliability or validity of scores they produce.

To evaluate the potential for behavior analytic assessments to inform medication effects for children with disabilities, we need to begin examining the scientific utility (i.e., reliability and validity) of scores produced from these assessments (Yoder et al., 2018). Particularly outside the stimulant literature, and especially in recent years, clinical studies incorporating any direct measures of medication effects are few and far between ([redacted for review], under review). Fewer still include standardized behavioral assessments appropriate for use in group design research (e.g., clinical trials) and with potential to isolate behavioral processes ([redacted for review], in preparation; see Grondhuis et al., 2019 and Handen et al., 2013 for exceptions). Among group design studies incorporating both indirect and direct measures of medication effects for children with disabilities, indirect measures have detected changes more often than direct measures (e.g., Aman et al., 1989; Snyder et al., 2002; Waxmonsky et al., 2010). This raises the question of whether the direct measures accurately reflected non-effects or lacked sufficient reliability or validity to detect them.

The purpose of the current study was to evaluate (a) reliability (i.e., temporal stability) and (b) construct validity of scores from two behavior analytic assessments—one designed to inform the value of positive reinforcement (progressive ratio assessment) and another designed to inform the value of negative reinforcement (demand assessment). To evaluate temporal stability, we conducted a series of generalizability and optimization studies to understand sources of score variance and identify conditions in which temporally stable estimates of behavior can be obtained. To inform construct validity, we calculated correlations between scores from each behavioral assessment and those from subscales of the Aberrant Behavior Checklist-Second Edition (ABC-2; Aman et al., 1985) to determine whether scores corresponded in the expected direction. We addressed the following research questions:

1. In a sample of children with disabilities who have been prescribed medication to address behavioral concerns, what proportion of the variance in (a) progressive ratio scores and (b) demand scores is due to true score variance versus facets of the measurement system (session, observer)? (c) Based on these variance component estimates, how many sessions of each assessment are needed to meet a minimum stability criterion ($g \geq .70$)?

2. For a subset of children who completed the same assessment battery 8 weeks after starting a new medication regimen, do variance structures change across time points?

3. Within each of two time points, are there associations between scores from each behavioral assessment and parent ratings of externalizing challenging behavior (i.e., Irritability and Hyperactivity/Noncompliance subscales of the ABC-2; Aman et al., 1985)?

Method

Participants

We recruited participants from a clinic in the [blinded]. To participate, children were required to (a) be 5–17 years of age, (b) have a disability and/or diagnosed psychiatric disorder, and (c) have been prescribed a psychotropic medication to address a behavioral concern. We aimed to recruit children who were prescribed a new medication or an increased dosage of a current medication but had not begun the new regimen. The prescribing physician (last author) shared recruitment flyers with families whose child met the above criteria. She provided input on the timing between behavioral assessments (see Procedures), but was not involved in the design, implementation, or evaluation of behavioral assessments throughout recruitment or data collection. We contacted families who expressed an interest in the study by phone to share more information on study goals and procedures. If the parent indicated further interest, we sent them a link to a consent form via email.

Twenty-three children participated in the study. Most participants were boys ($n = 17$; 73.9%) and the majority were White ($n = 14$; 60.9%). Eighteen of the 23 participants had an intellectual or developmental disability; the remaining five participants' primary diagnosis was ADHD—four of whom also had a speech/language impairment and one of whom also had an anxiety disorder. Ten participants had both an intellectual or developmental disability and a comorbid psychiatric disorder (e.g., ADHD, anxiety disorder). Parents reported all participants engaged in some form of challenging behavior, and more than one topography was reported for the majority (82.6%) of participants. While all participants had recently been prescribed a new medication or a new dosage of a current medication, 19 participants were already taking at least one psychotropic medication at the time of the initial assessment visit. Twenty-two participants returned for a second assessment visit approximately 8 weeks following their initial visit. We

confirmed that 18 of these children followed through with starting a new medication regimen shortly after the initial assessment visit. Additional details on participant characteristics and medication classes are presented in Table 1.

Setting

All behavioral assessments were completed in one of two clinic rooms on a university campus. Each room contained two tables and at least two chairs. One clinic room had a one-way mirror and an adjoining observation room with recording equipment. The other room was an empty office with no adjoining observation room. With few exceptions, two adults were in the room with the child during assessment sessions. One adult served as primary therapist (i.e., assessment implementer) and a second adult was present to support implementation as needed. For sessions conducted in the empty office, a third adult was present to work the video recorder.

Materials

Progressive ratio assessment materials included a large plastic container with a 2.1 x 2.1 cm hole cut into the lid, a set of two hundred 2-cm plastic cubes, a laminated picture of a stop sign, and a set of three to five snacks (e.g., fruit snacks, pretzels, m&ms). Demand assessment materials varied by participant, but often included a set of items to clean up (e.g., toys or blocks and a bin, crumpled pieces of paper and a trash can). We used Canon VIXIA video cameras on tripods to record all sessions. We used iPhones or iPods with a timed-event data collection application (Countee; Gavran & Hernandez, 2018) to collect primary data on child behavior and paper-pencil data collection forms to collect procedural fidelity data.

Procedures

After obtaining parent consent, we scheduled a 30-min parent phone interview to collect initial information on the child (i.e., demographics, prescribed medication, description of

problem behavior). To inform selection of edibles to use for progressive ratio sessions, we shared a menu of available snacks, and asked parents to indicate which ones their child preferred. If none were highly preferred, parents identified other preferred edibles and we procured them prior to the assessment visit. We also asked parents to identify non-preferred task demands for their child, which we used to inform demand assessment procedures. Finally, we scheduled the first assessment visit, which was typically within one week of the phone interview.

Parents brought their child to the university for two 1.5- to 2-hr assessment visits, which were scheduled approximately 8 weeks apart. We wanted to keep the time between assessment visits constant across participants, thus we selected a time span that was long enough for a therapeutic effect to be present at Time 2 across all medication classes likely to be prescribed (McVoy & Findling, 2017). Though the purpose of this study was not to evaluate medication effects directly, our goal was to time the assessment visits similarly to how future medication efficacy studies might be designed. At each visit, we completed three sessions for each of two assessments: progressive ratio and demand. We also completed four trials of a concurrent operant preference assessment, the results of which are not included in this paper. We randomized the sequence of sessions within three session blocks, with 5-min breaks between each block.

Progressive Ratio Assessment

Prior to each progressive ratio session, the therapist prompted the child to choose a snack from an array of three to five options. The therapist told the child they can put blocks in the bin to earn the snack, but can stop whenever they want. She modeled the target response (i.e., placed one block in the bin), prompted the child to practice the response, and delivered an edible following the response. She told the child that sometimes they might have to put a lot of blocks

in the bin to get the snack, and other times just a few. Finally, she told the child to say "all done"

or touch the stop sign if they wanted to stop, and modeled touching the stop sign. We used one of

two versions of scripted instructions depending on the child's communication skills. Across all

participants, we practiced the response and reward at least once before beginning the session.

During the session, the therapist reinforced the target response (i.e., putting one block in the bin)

according to a rapid additive progression (Fixed ratio [FR] 1, FR2, FR5, FR10, FR20, FR30,

FR40; Reed et al., 2009). We set a maximum FR value (40) and session duration (5 min) to

minimize the likelihood of problem behavior due to ratio strain (i.e., response requirements

becoming too large) and to ensure adequate time to conduct repeated sessions of each

assessment. Other researchers have taken a similar tactic in setting maximum session durations

and/or breakpoints (e.g., Chelonis et al., 2011; Paule et al., 1999; Reed et al., 2009). If problem

behavior occurred (9 of 135 sessions), the therapist provided a verbal reminder that the child

could keep working to earn [snack] or touch the stop card if they were all done. If problem

behavior continued following the verbal reminder (3 of 135 sessions), the therapist brought the

stop card to touch the child's hand (to avoid hand-over-hand prompting) and terminated the

session. Otherwise, the session ended when (a) the child indicated they wanted to stop, (b) 1 min

elapsed with no target responses, or (c) the child completed the final response requirement

(FR40). If the child had not met any of these criteria after 5 min elapsed, we ended the

assessment after delivery of the next reinforcer.

Demand Assessment

The demand assessment was modeled after the escape condition of a latency-based

functional analysis (Thomason-Sassi et al., 2011) as well as other demand assessments that use

latency to problem behavior as an index of task aversiveness (e.g., Call et al., 2016). During each

session, the therapist presented directives to complete a non-preferred task (e.g., cleaning up

toys, throwing away trash) using a three-step prompting sequence (i.e., verbal, model, physical).

Contingent on compliance, therapists provided brief praise followed by a subsequent task

directive. If the child did not comply within 5 s of the verbal prompt, the therapist provided a

model prompt; if the child did not comply within 5 s of the model prompt, the therapist provided

a physical prompt. Child requests to escape or change the activity were briefly acknowledged but

not reinforced (e.g., "maybe later, right now we're cleaning up"). Contingent on the first

occurrence of problem behavior, the therapist withdrew the demands (e.g., "Okay, we're all

done" while setting the task materials aside), at which point the session ended. The session ended

after 5 min if problem behavior did not occur. The 5-min session duration was informed by

previous studies using latency-based functional analyses (e.g., Hansen et al., 2019; Lambert et

al., 2017; Thomason-Sassi et al., 2011).

Measures

Parent-Completed Forms

During each assessment visit, parents completed a demographic and medication history

form and the ABC-2 (Aman et al., 1985). The demographic and medication history form

included items on basic child demographics and current and previous medications. The ABC-2

(Aman et al., 1985) is a 58-item behavior rating scale with five subscales (Irritability, Social

Withdrawal, Stereotypic Behavior, Hyperactivity/Noncompliance, and Inappropriate Speech)

that has an extensive history as an outcome measure in medication efficacy studies (Aman,

2015). When used by parents to rate the behavior of children and adolescents with

developmental disabilities, each of the ABC-2 subscales has high internal consistency

(Cronbach's α between .83 and .93) and high test-retest reliability over one month (Pearson's *r*

between .80 and .95; Freund & Reiss, 1991).

Direct Behavioral Assessments

From videos of progressive ratio sessions, we collected timed-event data (i.e., recorded the time at which each event occurred; Yoder et al., 2018) on frequencies of target responses and reinforcer deliveries. Though not a primary outcome measure for this assessment, we also collected data on frequencies of problem behavior (defined below). The target response was coded when the child placed a block in the bin. Reinforcer delivery was coded each time the therapist placed a small edible on a napkin or plate beside the child. For each progressive ratio session, we used the total number of responses and reinforcer deliveries to calculate the breakpoint (i.e., the response value of the last schedule completed before the child stopped responding, requested to stop, or reached the pre-defined session duration limit; Roane, 2008). We calculated target response rates by dividing the total frequency of responses by the session duration (min). We included a response rate measure to inform the temporal nature of response patterns within sessions. Unlike the breakpoint score, there was no maximum response rate.

From videos of demand sessions, we collected timed-event data on problem behavior, requests to escape the task or change activities, and child compliance with therapist task demands. Problem behavior was broadly defined to capture a range of reported topographies. We defined problem behavior to include aggression (i.e., forceful contact between the child's body and another person); disruption (i.e., forceful contact between the child's body and an object or surface; throwing or breaking objects; crying or screaming); and self-injury (i.e., forceful contact between the child's body and another part of their body or between their head and an external surface). These standard definitions applied across all participants. When parents described other forms of problem behavior that did not meet these definitions, we included additional

topographies in the problem behavior definition on an individual basis. These additional

topographies included active noncompliance (i.e., overt verbal or gestural refusals to follow

adult-given directives, such as yelling "No!" or pushing away materials); inappropriate language

(e.g., cursing, making verbal threats of harm); and elopement (i.e., attempts to leave the

assessment room by [a] placing a hand on the door knob or [b] moving at least three steps away

from the therapist or session materials towards the door). We defined requests as any vocal or

gestural bid to escape the activity or engage in an activity that was incompatible with completing

the task demand, excluding behaviors that would meet the above definitions for problem

behavior. Example requests included using a communication device to request iPad, asking to

take a walk or see the parent, and pointing to materials from another assessment activity. We

defined compliance as the initiation of task completion within 10 s of a therapist verbal prompt

(i.e., if the child complied following the verbal or model prompt). Latencies to problem behavior

were coded from the timed-event data files, as were latencies to the first instance of a request.

For sessions in which a request did not occur but problem behavior did (thus ending the session),

the latency to problem behavior was entered for both latency measures. Thus, the latency to

request score represented the latency to request *or problem behavior*, whichever occurred first.

We conceptualized this score as a broader measure of tolerance, indicating the point at which the

child attempted to initiate a change in activities, either by requesting it outright or engaging in

problem behavior. We calculated a percentage of compliance by dividing the number of

occurrences of compliance by the total number of task demands and multiplying by 100.

Therapist and Coder Training

Graduate research assistants in a department of Special Education implemented and

coded all assessment sessions. Before serving as therapist in study sessions, research assistants

completed the following training activities. First, they attended a 60-min training meeting that
included didactic instruction on each assessment and role-play with feedback. Therapists then
watched example videos of the assessments and practiced role-playing with another research
team member. Finally, each therapist practiced implementing each assessment with the third
author until they correctly implemented all procedures. Before collecting study data, research
assistants attended a 45- to 60-min didactic training on data collection procedures and were
required to reach 90% agreement with master codes across three consecutive sessions of each
assessment for both child behavior and procedural fidelity variables.

Procedural Fidelity

Data collectors used paper-pencil forms to collect procedural fidelity data on therapist
implementation across all sessions. For progressive ratio sessions, they indicated the presence or
absence of four pre-session components (i.e., correct materials present, read script with model,
practiced response and reward, solicited and answered questions). Then, for each schedule
requirement completed in the session, they indicated whether each of the following procedures
were correctly implemented (Yes, No, or Not applicable): followed prompting rules, minimized
attention, delivered reward upon schedule completion, blocked attempts to put more than one
block in the bin at a time, redirected problem behavior, and followed session termination
procedures. For each session, we calculated a percentage of correct implementation by dividing
the total number of Yes tallies by the sum of Yes and No tallies and multiplying by 100. Mean
fidelity for progressive ratio sessions was 98.3% (range, 93.4%–100%) for pre-session
components and 96.1% (range, 85.7%–100%) for all other within-session components.

For demand sessions, coders scored the presence or absence of one pre-session
component (i.e., correct materials present) and the session termination criteria (i.e., ended session

within 2 s of the first occurrence of problem behavior or when 5 min elapsed). In addition,

sessions were divided into 10-s intervals, and for each interval, coders scored whether the

following three procedural components were correctly implemented: presented demand-related

prompts every 5 s when the child was not complying, used the appropriate prompting sequence,

and delivered brief praise following compliance. For intervals scored as containing an error,

coders indicated which of these three procedural components was implemented incorrectly.

Therapists had correct materials present in 100% of sessions, and correctly terminated the

session in 96.3% of sessions. The mean percentage of intervals with correct implementation was

97.6% (range, 57.1%–100%). The most common error was more than 5 s elapsing between

therapist demand deliveries in the absence of child compliance. Sessions with low fidelity were

those that were brief, resulting in relatively few opportunities to score correct implementation.

Inter-Observer Agreement

All sessions were coded independently by two trained data collectors. For formative

assessment of inter-observer agreement (IOA), we calculated total agreement percentages (i.e.,

[smaller/larger]*100) and monitored agreement percentages produced from the Countee software

(i.e., 10-s interval-by-interval agreement). For progressive ratio sessions, mean agreement was

99.3% (range, 83.3%–100%) for target responses, 97.5% (range, 50.0%–100%) for reinforcer

delivery, and 99.5% (range, 83.3%–100%) for problem behavior. For demand sessions, mean

agreement was 99.7% (range, 87.5%–100%) for latency to problem behavior, 92.8% (range,

3.3%–100%) for latency to requests, and 93.3% (range, 66.7%–100%) for compliance. Low

minimum agreement scores for latency to requests were identified for sessions in which

observers disagreed on whether a request occurred very early in a session (producing two very

different latencies). We calculated point-by-point agreement on procedural fidelity of

progressive ratio sessions as the number of agreements on each fidelity code (Yes, No, N/A)

divided by the number of agreements plus disagreements, multiplied by 100. We calculated a

percentage of intervals with agreement on procedural fidelity codes for demand sessions

(intervals with errors had to have the same error type coded to count as an agreement). Mean

agreement was 97.3% (range, 66.7%–100%) for progressive ratio sessions and 96.2% (range,

50.0%–100%) for demand sessions. Sessions with low agreement often reflected single

disagreements (or single intervals with disagreement) among few opportunities to agree.

Data Analysis

To address the first research question, we conducted a series of fully crossed, 3-facet

(random effects) generalizability studies using EduG software (Swiss Society for Research in

Education Working Group, 2012). EduG uses Type III mean squares from analyses of variance

(ANOVAs) to calculate a series of variance component estimates. Our 3-facet generalizability

studies focused on variance due to Person (i.e., true score variance) and variance due to two

facets of the measurement system (i.e., Session and Observer). We were particularly interested in

isolating the percentages of variance accounted for by Person, which represents true score

variance; Person x Session, which represents the extent to which participant score rankings vary

by session; and Person x Observer, which represents the extent to which participant score

rankings vary by observer.

We also calculated absolute *g* coefficients (Shavelson & Webb, 2006) for each behavior

assessment score using EduG. A *g* coefficient is a type of intra-class correlation that represents

the amount of variance due to Person (true score) divided by the total observed score variance

(true score + measurement error; Yoder et al., 2018). Essentially, these *g* coefficients indicated

whether the number of sessions (3) and observers (2) used for this study were sufficient to

produce adequately stable scores ($g > .70$; Berk, 1979; Bloch & Norman, 2012; Mitchell, 1979)

given the observed variance. Formulas for calculating variance component estimates and g

coefficients are accessible via the following link: [redacted]. We then used EduG to conduct

optimization studies, which use the variance component estimates from the generalizability

studies to estimate projected g coefficients for different measurement structures (e.g., varying the

number of sessions or observers from what was used for the existing data set; Yoder et al., 2018).

Our primary goal in conducting optimization studies was to determine the minimum number of

sessions required to produce adequately stable scores for future studies.

To address Research Questions 1–2, we conducted separate generalizability and

optimization studies using (a) all data from the initial assessment visit (23 Persons x 3 Sessions x

2 Observers) and (b) data from 18 participants who started a new medication regimen before

returning 8 weeks later to complete the same assessment battery (18 Persons x 3 Sessions x 2

Observers). To address the third research question, we calculated Pearson Product-Moment

correlations between each behavior assessment score (using averages across the three sessions)

and subscale scores from the ABC-2 within each time point. We predicted negative associations

between each behavior assessment score and each of two subscales related to externalizing

challenging behavior (i.e., Irritability and Hyperactivity/Noncompliance). That is, we expected

the children who were motivated to engage in an adult-defined target response to earn positive

reinforcers (evidenced by high breakpoints or response rates during progressive ratio sessions)

would be rated by parents as exhibiting lower levels of irritability, hyperactivity, and/or

noncompliance outside the assessment context. Similarly, we expected the children who were

able to tolerate and comply with non-preferred task demands (evidenced by long latencies to

problem behavior or requests, and high percentages of compliance during demand sessions)

would be rated by parents as exhibiting lower levels of irritability, hyperactivity, and/or

noncompliance outside the assessment context. Because separate sets of correlations were

calculated and tested at each time point and were not intended to inform changes following new

medication regimens, we included data from all completed assessments at each time point in the

correlational analysis. We used SPSS version 26 (IBM SPSS Statistics for Macintosh) to test the

statistical significance of these correlations (two-tailed significance tests against alpha values of

.05 and .01).

Results

Table 2 presents descriptive statistics for behavioral assessment scores for all participants

who completed the initial assessment visit, all participants who completed a second assessment

visit, and the subset of participants within each time point for whom one or more medication

change was made. Frequency distributions of behavioral assessment scores by time point are

presented in Figure 1. Means, standard deviations, and frequency distributions suggest adequate

among-participant variance for progressive ratio scores but a ceiling effect for demand scores,

particularly latency to problem behavior. During initial assessment visits, 16 of 23 participants

did not engage in problem behavior across demand sessions (i.e., produced a maximum latency

score of 300 s). To explore the potential for systematic differences in behavioral assessment

scores by medication class, we identified scores representing the most common medication class

across time points (i.e., stimulants; see closed triangles in Figure 1). We saw no evidence of

systematic differences.

Results of generalizability studies on scores from all participants at the initial assessment

visit are shown in Table 3. For progressive ratio scores, the percentage of variance accounted for

by Person (true score variance) exceeded 60% for breakpoint and 80% for response rate. The

Person x Session interaction accounted for more than 30% of variance in breakpoint scores, whereas this interaction accounted for roughly 15% of variance in response rate scores. The Person x Observer interaction accounted for 0% of variance across scores. These patterns suggest that participants would be ranked differently depending on which session was selected, particularly for break point scores. This indicates a need to average scores across more than one session to obtain adequately stable estimates. Using data from all three sessions and two observers, the absolute g coefficient for both progressive ratio scores exceeded .80.

Percentages of variance accounted for by Person (true score variance) varied among demand assessment scores at Time 1 (34.9%–55.7%), and were lower than estimates for progressive ratio scores. Again, the Person x Session interaction accounted for substantial variance in demand scores relative to the Person x Observer interaction, suggesting a need to average scores across multiple sessions to achieve adequate stability. Using data from all three sessions and two observers, absolute g coefficients ranged from .62 (percentage compliance) to .80 (latency to problem behavior). Across all behavioral assessment scores, main effects of Session and Observer as well as the Session x Observer interaction contributed trivial variance. Random error (i.e., Person x Session x Observer) also contributed trivial variance with one exception (i.e., Latency to request at Time 2 [15.0%]).

Results of optimization studies are shown in Figure 2, which depicts projected g coefficients across varying numbers of sessions for progressive ratio scores (top graph) and demand scores (bottom graph). These data reflect scores from single observers, given the near-zero variance component estimates associated with the observer facet. Results suggested reliable estimates of response rate can be obtained from a single progressive ratio session, whereas reliable estimates of breakpoint require at least two. A minimum of two demand sessions are

needed to obtain stable estimates of latency to problem behavior, whereas four or five sessions

are needed to obtain stable estimates of latency to request and percentage of compliance.

Results of generalizability studies on scores from the set of participants who started a

new medication regimen are presented in Table 4. Across scores, percentages of variance

accounted for by Person (true score variance) increased from Time 1 to Time 2, and percentages

of variance accounted for by the Person x Session interaction decreased. These changes in

variance component estimates resulted in higher g coefficients at Time 2 relative to Time 1. With

one exception (latency to request at Time 2), variance accounted for by the Person x Observer

interaction remained zero or trivial. Based on the increased stability of scores at Time 2, we did

not conduct follow up optimization studies, as the number of sessions required to achieve

stability would have been equal to or fewer than the number identified for the initial data set.

Pearson product-moment correlations between each behavioral assessment score and

ABC subscale score are shown in Table 5. Three behavioral assessment scores were significantly

and negatively associated with parent ratings of Irritability: breakpoint (Time 1 only), response

rate (Time 1 only), and percentage compliance. Four behavioral assessment scores were

significantly associated with parent ratings of Hyperactivity/Noncompliance: breakpoint,

response rate, latency to problem behavior (Time 1 only), and percentage of compliance. Though

not predicted, two behavioral assessment scores were also significantly and negatively associated

with parent ratings of Stereotypy (response rate [Time 1 only], percentage compliance), and one

score (breakpoint, Time 2 only) was significantly and negatively associated with parent ratings

of Inappropriate Speech. None of the behavioral assessment scores were significantly associated

with Social Withdrawal subscale scores.

Discussion

The goal of this study was to explore the potential for each of two behavior analytic assessments to inform whether and how psychotropic medication impacts behavioral processes (i.e., positive and negative reinforcement). In this initial evaluation, we focused on temporal stability and construct validity of scores produced from each assessment. Results highlighted three main findings that inform future research applying these assessments for purposes of medication evaluation. First, a ceiling effect was observed for demand assessment scores, and was particularly pronounced for latency to problem behavior. Though the 5-min session maximum was informed by previous research on functional analysis (e.g., Hansen et al., 2019; Thomason-Sassi et al., 2011; Wallace & Iwata, 1999), our demand sessions did not occur in the context of a functional analysis, but were interspersed with other assessment types. In addition, though we selected task demands based on what parents reported would be likely to evoke problem behavior, we did not collect any preliminary data to confirm this pattern. In future studies, we recommend increasing the maximum duration of demand sessions from 5 to 10 min. Increasing the maximum session duration is expected to improve the degree to which scores would be sensitive to change after beginning a new medication regimen.

Second, with only one exception, scores from single sessions did not have the temporal stability needed to reliably differentiate among participants. However, the variance estimates obtained from the current data set suggest that averaging across as few as two sessions would be sufficient to obtain adequately stable estimates ($g > .70$) of progressive ratio scores without compromising their content validity (e.g., reinforcers losing value as a result of conducting too many sessions). Response rates showed more stability relative to breakpoints, which we attribute to the increased range of possible values relative to the number of possible breakpoint scores. Results also suggest averaging across as few as two sessions could be sufficient to obtain

adequately stable estimates of demand scores if latency to problem behavior is the primary

outcome measure, whereas several more sessions would be needed to obtain reliable estimates of

latencies to requests and percentages of compliance. We suspect extending session durations

from 5 to 10 min would increase the stability of demand scores such that fewer sessions would

be needed to reach reliability thresholds. Variance estimates for the observer facet were zero or

near-zero across almost all scores. The single exception was the latency to request score, for

which percentages of agreement at the session level were also lower relative to other scores.

These results suggest that averaging scores from multiple observers is not necessary as long as

formative point-by-point interobserver agreement estimates are within acceptable ranges.

Third, significant negative correlations between scores from behavior analytic

assessments and ABC-2 subscale scores related to externalizing challenging behavior offer initial

evidence of construct validity. In particular, breakpoint, response rate, and percentage of

compliance were all significantly and negatively correlated with Irritability and

Hyperactivity/Noncompliance subscale scores at one or both time points. We should point out,

however, that continued evaluation of construct validity for these behavioral assessments is

critical. We selected the ABC-2 based on its widespread use as a behavioral outcome measure in

medication evaluations for children with disabilities. Other types of indirect assessments, such as

caregiver ratings of problem behavior function, might be expected to align more closely with

direct measures of positive and negative reinforcer value and provide stronger evidence of

construct validity. It is also possible that associations between direct assessment scores and

ABC-2 subscale scores vary by behavior function. Including multiple and varied assessment

measures in future studies would offer further opportunities to evaluate construct validity of

scores from each behavioral assessment.

Taken together, our results support the scientific utility of progressive ratio scores, as evidenced by the among-participant variance observed (Figure 1), their high temporal stability (Tables 3–4), and significant negative correlations with parent ratings of challenging behavior (Table 5). On the other hand, the ceiling effect observed for demand scores, as well as the low temporal stability of percentage of compliance, raises questions about whether the observed negative linear relations between demand scores and parent ratings of challenging behavior are replicable. Future studies are needed to evaluate temporal stability and construct validity for demand sessions with longer maximum durations. In addition to extending session duration, researchers might also consider the use of aggregate variables for demand assessments (i.e., combining multiple component scores by averaging or summing them). Particularly in the context of group design research, aggregating scores from one assessment has potential to improve both temporal stability and construct validity (Yoder et al., 2018).

Our findings should be interpreted with the following limitations in mind. First, participants in this study represent a heterogeneous sample, with respect to both diagnoses and medications prescribed. We do not know if results would replicate for more homogenous samples (e.g., children with autism as only diagnosis) or those prescribed medications that were minimally represented in this sample (e.g., atypical antipsychotics). Second, due to the variety of medication changes made after the first assessment visit, our analyses focused on temporal stability and construct validity within time point, rather than evaluating changes in behavior (or behavior ratings) between time points. Additional studies are needed that focus on a single class of medication to address sensitivity to change associated with each behavioral assessment. Third, we selected progressive ratio and demand assessments as potential indicators of positive and negative reinforcer value, respectively, but we did not experimentally evaluate whether edibles

used in the progressive ratio sessions indeed functioned as positive reinforcers; nor did we experimentally confirm that participants' problem behavior was maintained by negative reinforcement. In the future, brief control conditions might be added to the assessment battery to better isolate each behavioral process. For example, response rates from progressive ratio sessions might be compared to those observed when no edible rewards are delivered contingent on the target response. Or, latencies to problem behavior from demand sessions might be compared to those observed in free play sessions in which no demands are presented. Finally, we did not control for the potential influence of side effects on behavioral assessment scores. For example, decreased appetite is a common side effect of stimulant medication that could influence the value of edible reinforcers used during progressive ratio sessions. We did not find evidence of progressive ratio scores being systematically lower in the stimulant subgroup relative to other participants (see Figure 1). But in future research, especially studies focused on a single medication class, procedural adaptations to behavioral assessments may be warranted to clearly distinguish therapeutic effects of medication from common side effects.

The existing literature on behavior analytic outcome measures for medication effects is limited with respect to both quantity and quality ([redacted for review], in preparation; Courtemanche et al., 2011; Napolitano et al., 1999). The current study represents the first evaluation of measurement properties of two behavior analytic assessments with potential to inform how medication might impact positive and negative reinforcement processes for children with disabilities and behavioral concerns. Results of this initial investigation inform conditions in which scores from each assessment are likely to produce temporally stable estimates of behavior and offer preliminary evidence of construct validity. Taken together, our results suggest break point and response rate scores from progressive ratio assessments may be scientifically useful in

evaluating change in positive reinforcer value following medication. More research is needed to

evaluate sensitivity to change in scores from these and other behavior analytic assessments for

particular classes of medication. In addition, similar evaluations of temporal stability and

construct validity are needed for behavior analytic assessments designed to inform other

behavioral processes (e.g., stimulus control, sensitivity to parameters of reinforcement,

impulsivity), for which the current study might serve as a model.

References

[Redacted for review]

American Academy of Child and Adolescent Psychiatry. (2012). *A guide for community child serving agencies on psychotropic medications for children and adolescents.* http://www.aacap.org/app_themes/aacap/docs/press/guide_for_community_child_serving _agencies_on_psychotropic_medications_for_children_and_adolescents_2012.pdf

Aman, M. G. (2015). *Annotated bibliography on the Aberrant Behavior Checklist (ABC).* The Nisonger Center. https://psychmed.osu.edu/wp-content/uploads/2017/02/ABC_Annotated_Bibliography_08-28-2015.pdf

Aman, M. G., Novotny, S., Samango-Sprouse, C., Lecavalier, L., Leonard, E., Gadow, K., King, B. H., Pearson, D. A., Gernsbacher, M. A., & Chez, M. (2004). Outcome measures for clinical drug trials in autism. *CNS Spectrums, 9*(1), 36–47. https://doi.org/d2fd

Aman, M. G., Singh, N. N., Stewart, A. W., & Field, C. J. (1985). The Aberrant Behavior Checklist: A behavior rating scale for the assessment of treatment effects. *American Journal of Mental Deficiency, 89*(5), 485–491.

Aman, M. G., Teehan, C. J., White, A. J., Turbott, S. H., & Vaithianathan, C. (1989). Haloperidol treatment with chronically medicated residents: Dose effects on clinical behavior and reinforcement contingencies. *American Journal on Mental Retardation, 93*(4), 452–460.

Angold, A. Erkanli, A., Egger, H. L., & Costello, E. J. (2000). Stimulant treatment for children: A community perspective. *Journal of the American Academy of Child and Adolescent Psychiatry, 39*(8), 975–984. https://doi.org/10.1097/00004583-200008000-00009

Berk, R. A. (1979). Generalizability of behavioral observations: A clarification of interobserver

agreement and interobserver reliability. *American Journal of Mental Deficiency, 83*, 460–
472.

Bloch, R., & Norman, G. (2012). Generalizability theory for the perplexed: A practical
introduction and guide: AMEE Guide No. 68, *Medical Teacher, 34*(11), 960–992.
https://doi.org/gf638v

Call, N. A., Pabico, R. S., & Lomas, J. E. (2009). Use of latency to problem behavior to evaluate
demands for inclusion in functional analyses. *Journal of Applied Behavior Analysis, 42*,
723–728. https://doi.org/10.1901/jaba.2009.42-723

Call, N. A., Miller, S. J., Mintz, J. C., Mevers, J. L., Scheithauer, M. C., Eshelman, J. E., &
Beavers, G. A. (2016). Use of a latency-based demand assessment to identify potential
demands for functional analyses. *Journal of Applied Behavior Analysis, 49*, 900–914.
https://doi.org/10.1002/jaba.341

Carlson, J. S. (2019). Introduction to school psychopharmacology. In J. S. Carlson & J. A.
Barterian (Eds.), *School psychopharmacology: Translating research into practice* (pp. 1–
11). Springer.

Carlson, G., Pokrzywinski, J., Uran, K., & Valdovinos, M. (2012). The use of reinforcer
assessments in evaluating psychotropic medication effects. *Journal of Developmental and
Physical Disabilities, 24*(5), 515–528. https://doi.org/10.1007/s10882-012-9282-4

Chelonis, J. J., Johnson, T. A., Ferguson, S. A., Berry, K. J., Kubacak, B., Edwards, M. C., &
Paule, M. G. (2011). Effect of methylphenidate on motivation in children with attention-
deficit/hyperactivity disorder. *Experimental and Clinical Psychopharmacology, 19*, 145–
153. https://doi.org/10.1037/a0022794

Cox, A. D., & Virues-Ortega, J. (2016). A review of how psychotropic medication can affect the

motivation of challenging behavior. *International Journal of Developmental Disabilities,*

62(3), 192–199. https://doi.org/10.1080/20473869.2016.1175157

Crosland, K. A., Zarcone, J. R., Lindauer, S. E., Valdovinos, M. G., Zarcone, T. J., Hellings, J.

A., & Schroeder, S. R. (2003). Use of functional analysis methodology in evaluation of

medication effects. *Journal of Autism and Developmental Disorders, 33*(3), 271–279.

https://doi.org/10.1023/A:1024402500425

Danov, S. E., Tervo, R., Meyers, S., & Symons, F. J. (2012). Using functional analysis

methodology to evaluate effects of an atypical antipsychotic on severe problem behavior.

Journal of Mental Health Research in Intellectual Disabilities, 5(3–4), 286–308.

https://doi.org/d2ff

Dicesare, A., McAdam, D. B., Toner, A., & Varnell, J. (2005). The effects of methylphenidate

on a functional analysis of disruptive behavior. A replication and extension. *Journal of

Applied Behavior Analysis, 38*(1), 125–128. https://doi.org/10.1901/jaba.2005.155-03

Freund, L. S., & Reiss, A. L. (1991). Rating problem behaviors in outpatients with mental

retardation: Use of the Aberrant Behavior Checklist. *Research in Developmental

Disabilities*, *12*(4), 435–451. https://doi.org/10.1016/0891-4222(91)90037-s

Gavran, D. P., & Hernandez, V. (2018). *Countee* (Version 2.0.0) [Mobile app]. App Store.

https://apps.apple.com/us/app/countee-data-collection-system/id982547332?ls=1

Grondhuis, S. N., Farmer, C. A., Arnold, L. E., Gadow, K. D., Findling, R. L., Molina, B. S. G.,

Kolko, D. J., Buchan-Page, K. A., Rice, R. R., Butter, E. M., & Aman, M. G. (2019).

Standardized observation analogue procedure in the treatment of severe childhood

aggression study. *Journal of Child and Adolescent Psychopharmacology, 30*(1), 48–54.

https://doi.org/d7nx

Guy, W. (1976). Clinical Global Impressions. In W. Guy (Ed.), *ECDEU Assessment Manual for*

    Psychopharmacology, (Revised, pp. 217–221). National Institute of Mental Health.

Handen, B. L., Johnson, C. R., Butter, E. M., Lecavalier, L., Scahill, L., Aman, M. G.,

    McDougle, C. J., Arnold, L. E., Swiezy, N. B., Sukhodolsky, D. G., Mulick, J. A., White, S.

    W., Bearss, K., Hollway, J. A., Stigler, K. A., Dziura, J., Yu, S., Sacco, K., & Vitiello, B.

    (2013). Use of a direct observational measure in a trial of risperidone and parent training in

    children with pervasive developmental disorders. *Journal of Developmental and Physical*

    Disabilities, *25*(3), 355–371. https://doi.org/10.1007/s10882-012-9316-y

Hansen, B. D., Sabey, C. V., Rich, M., Marr, D., Robins, N., & Barnett, S. (2019). Latency-

    based functional analysis in schools: Correspondence and differences across

    environments. *Behavioral Interventions, 34*, 366–376. https://doi.org/10.1002/bin.1674

Kollins, S. H., Ehrhardt, K., & Poling, A. (2000). Clinical drug assessment. In A. Poling & T.

    Byrne (Eds.), *Introduction to behavioral pharmacology* (pp. 191–218). Context Press.

Lambert, J. M., Staubitz, J. E., Roane, J. T., Houchins-Juarez, N. J., Juarez, A. P., Sanders, K. B.,

    & Warren, Z. E. (2017). Outcome summaries of latency-based functional analyses

    conducted in hospital inpatient units. *Journal of Applied Behavior Analysis, 50*, 487–494.

    https://doi.org/10.1002/jaba.399

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. https://doi.org/10.1901/jaba.2003.36-407

Lloyd, B. P., Torelli, J. N., & Symons, F. J. (2016). Issues in integrating psychotropic and

    intensive behavioral interventions for students with emotional and behavioral challenges

    in schools. *Journal of Emotional and Behavioral Disorders, 24*(3), 148–158.

https://doi.org/f83jbb

Mandell, D. S., Morales, K. H., Marcus, S. C., Stahmer, A. C., Doshi, J., & Polsky, D. E. (2008).

> Psychotropic medication use among Medicaid-enrolled children with autism spectrum

> disorders. *Pediatrics, 121*(3), e441–e448. https://doi.org/10.1542/peds.2007-0984

Marcus, R. N., Owen, R., Kamen, L., Manos, G., McQuade, R. D., Carson, W. H., & Aman, M.

> G. (2009). A placebo-controlled, fixed-dose study of aripiprazole in children and

> adolescents with irritability associated with autistic disorder. *Journal of the American*

> *Academy of Child and Adolescent Psychiatry, 48*(11), 1110–1119. https://doi.org/c9jqz3

Mattison, R. E., Rundberg-Rivera, V., & Michel, C. (2014). Psychotropic medication

> characteristics for special education students with emotional and/or behavioral

> disorders. *Journal of Child and Adolescent Psychopharmacology, 24*(6), 347–353.

> https://doi.org/10.1089/cap.2013.0073

McCracken, J. T., McGough, J., Shah, B., Cronin, P., Hong, D., Aman, M. G., Arnold, L. E.,

> Lindsey, R., Nash, P., Hollway, J., McDougle, C. J., Posey, D., Swiezy, N., Kohn, A.,

> Scahill, L., Martin, A., Koenig, K., Volkmar, F., Carroll, D…McMahon, D. (2002).

> Risperidone in children with autism and serious behavioral problems. *The New England*

> *Journal of Medicine, 347*, 314–321. https://doi.org/10.1056/NEJMoa013171

McLaren, J. L., Barnett, E. R., Concepcion Zayas, M. T., Lichtenstein, J., Acquilano, S. C.,

> Schwartz, L. M., Woloshin, S., & Drake, R. E. (2018). Psychotropic medications for highly

> vulnerable children. *Expert Opinion on Pharmacotherapy*, *19*(6), 547–560.

> https://doi.org/ggk73n

McVoy, M., & Findling, R. L. (Eds.). (2017). *Clinical manual of child and adolescent*

> *psychopharmacology* (3rd ed.). American Psychiatric Association.

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected

in observational studies. *Psychological Bulletin, 86*, 376–390. https://doi.org/b6ch5p

Northup, J., Fusilier, I., Swanson, V., Roane, H., & Borrero, J. (1997). An evaluation of

methylphenidate as a potential establishing operation for some common classroom

reinforcers. *Journal of Applied Behavior Analysis, 30*(4), 615–625. https://doi.org/fn89bf

Northup, J., & Gulley, V. (2001). Some contributions of functional analysis to the assessment of

behaviors associated with attention deficit hyperactivity disorder and the effects of

stimulant medication. *School Psychology Review, 30*(2), 227-38.

Olfson, M., Druss, B. G., & Marcus, S. C. (2015). Trends in mental health care among children

and adolescents. *New England Journal of Medicine, 372*(21), 2029–2038.

https://doi.org/ggj2s3

Paule, M. G., Chelonis, J. J., Buffalo, E. A., Blake, D. J., & Casey, P. H. (1999). Operant test

batter performance in children: Correlation with IQ. *Neurotoxicology and Teratology, 21*,

223–230. https://doi.org/10.1016/s0892-0362(98)00045-2

Reed, D. D., Luiselli, J. K., Magnuson, J. D., Fillers, S., Vieira, S., & Rue, H. C. (2009). A

comparison between traditional economical and demand curve analyses of relative

reinforcer efficacy in the validation of preference assessment predictions. *Developmental

Neurorehabilitation*, *12*(3), 164–169. https://doi.org/10.1080/17518420902858983

Roane, H. S. (2008). On the applied use of progressive-ratio schedules of reinforcement. *Journal

of Applied Behavior Analysis, 41*(2), 155–161. https://doi.org/10.1901/jaba.2008.41-155

Rush, A. J., & Frances, A. (Eds.). (2000). Expert consensus guidelines series: Treatment of

psychiatric and behavioral problems in mental retardation. *American Journal on Mental

Retardation, 105*(3), 159–228.

Ryan, J. B., Reid, R., Gallagher, K., & Ellis, C. (2008). Prevalence rates of psychotropic

    medications for students placed in residential facilities. *Behavioral Disorders, 33*(2), 99–

    107. https://doi.org/10.1177/019874290803300204

Shavelson, R. J., & Webb, N. M. (2006). Generalizability theory. In J. L. Green, G. Camilli, P.

    B. Elmore & American Educational Research Association (Eds.), *Handbook of*

    complementary methods in education research. (3rd ed., pp. 309–322). Lawrence

    Erlbaum Associates.

Snyder, R., Turgay, A., Aman, M., Binder, C., Fisman, S., & Carroll, A. (2002). Effects of

    risperidone on conduct and disruptive behavior disorders in children with subaverage

    IQs. *Journal of the American Academy of Child & Adolescent Psychiatry*, *41*(9), 1026–

    1036. https://doi.org/10.1097/00004583-200209000-00002

Spencer, D., Marshall, J., Post, B., Kulakodlu, M., Newschaffer, C., Dennen, T., Azocar, F., &.

    Jain, A. (2013). Psychotropic medication use and polypharmacy in children with autism

    spectrum disorders. *Pediatrics, 132*(5), 833–840. https://doi.org/10.1542/peds.2012-3774

Swiss Society for Research in Education Working Group. (2012). *EduG* (Version 6.1)

    [Computer software]. IRDP. https://www.irdp.ch/institut/english-program-1968.html

Thomason-Sassi, J., Iwata, B. A., Neidert, P. L., & Roscoe, E. M. (2011). Response latency as an

    index of response strength during functional analyses of problem behavior. *Journal of*

    Applied Behavior Analysis, 44(1), 51−67. https://doi.org/10.1901/jaba.2011.44-51

Thompson, T., Moore, T., & Symons, F. (2007). Psychotherapeutic medications and positive

    behavior support. In S. L. Odom, R. H., Horner, M. E. Snell, & J. Blacher (Eds.),

    Handbook of developmental disabilities (pp. 501−527). Guilford Press.

Valdovinos, M. G., & Kennedy, C. H. (2004). A behavior-analytic conceptualization of the side

effects of psychotropic medication. *The Behavior Analyst, 27*(2), 231–238.

https://doi.org/d2fc

Vitiello, B. (2017). Developmental aspects of pediatric psychopharmacology. In M. McVoy & R.

L. Findling (Eds.), *Clinical manual of child and adolescent psychopharmacology* (3rd ed,

pp. 1–28). American Psychiatric Association.

Volpe, R. J., Heick, P. F., & Guerasko-Moore, D. (2005). An agile behavioral model for

monitoring the effects of stimulant medication in school settings. *Psychology in the

Schools, 42*(5), 509–523. https://doi.org/10.1002/pits.20088

Volpe, R. J., Daniels, B., & Sakai, C. (2019). School-based medication evaluations: Implications

for school personnel and physicians. In J. S. Carlson & J. A. Barterian (Eds.), *School

psychopharmacology: Translating research into practice* (pp. 213–230). Springer.

Wallace, M. D., & Iwata, B. A. (1999). Effects of session duration on functional analysis

outcomes. *Journal of Applied Behavior Analysis, 32*, 175–183. https://doi.org/bhcc7m

Waxmonsky, J. G., Waschbusch, D. A., Pelham, W. E., Draganac-Cardona, L., Rotella, B., &

Ryan, L. (2010). Effects of atomoxetine with and without behavior therapy on the school

and home functioning of children with attention-deficit/hyperactivity disorder. *The

Journal of Clinical Psychiatry, 71*(11), 1535–1551. https://doi.org/cjwd64

Weeden, M., Ehrhardt, K., & Poling, A. (2009). Conspicuous by their absence: Studies

comparing and combining risperidone and applied behavior analysis to reduce

challenging behavior in children with autism. *Research in Autism Spectrum

Disorders, 3*(4), 905–912. https://doi.org/10.1016/j.rasd.2009.06.004

Yoder, P. J., Symons, F. J., & Lloyd, B. P. (2018). *Observational measurement of behavior* (2nd

ed.). Paul H. Brookes Publishing Co.

Zarcone, J. R., Lindauer, S. E., Morse, P. S., Crosland, K. A., Valdovinos, M. G., McKerchar, T. L., Reese, M. R., Hellings, J. A., & Schroeder, S. R. (2004). Effects of risperidone on destructive behavior of persons with developmental disabilities: III. Functional analysis. *American Journal on Mental Retardation, 109*(4), 310–321. https://doi.org/cps877

Zarcone, J., Napolitano, D., & Valdovinos, M. (2008). Measurement of problem behaviour during medication evaluations. *Journal of Intellectual Disability Research, 52*(12), 1015–1028. https://doi.org/10.1111/j.1365-2788.2008.01109.x

Zito, J. M., Derivan, A. T., Kratochvil, C. J., Safer, D. J., Fegert, J. M., & Greenhill, L. L. (2008). Off-label psychopharmacologic prescribing for children: History supports close clinical monitoring. *Child and Adolescent Psychiatry and Mental Health, 2*(1), 24. https://doi.org/10.1186/1753-2000-2-24

Figure1
Figure 1

Frequency Distributions of Behavioral Assessment Scores by Time Point

Progressive Ratio Assessment
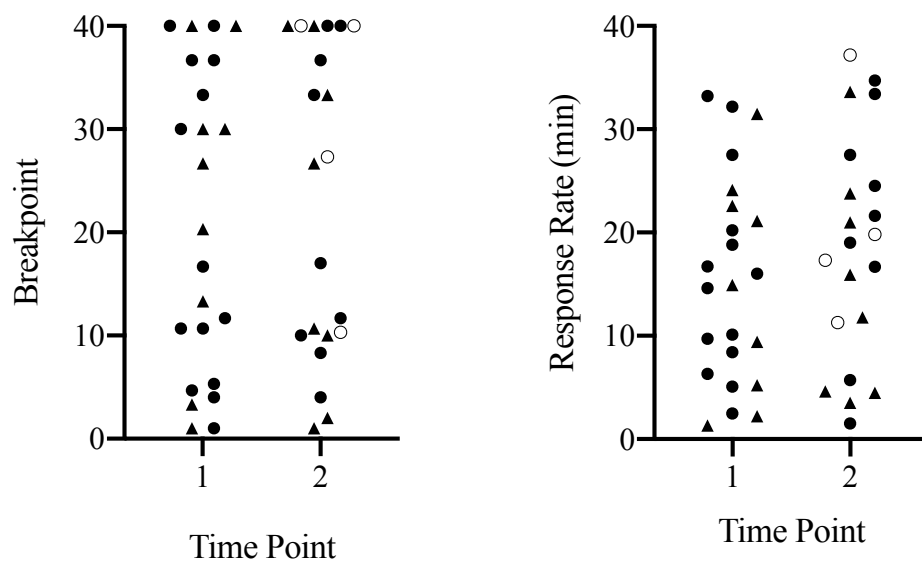
Demand Assessment
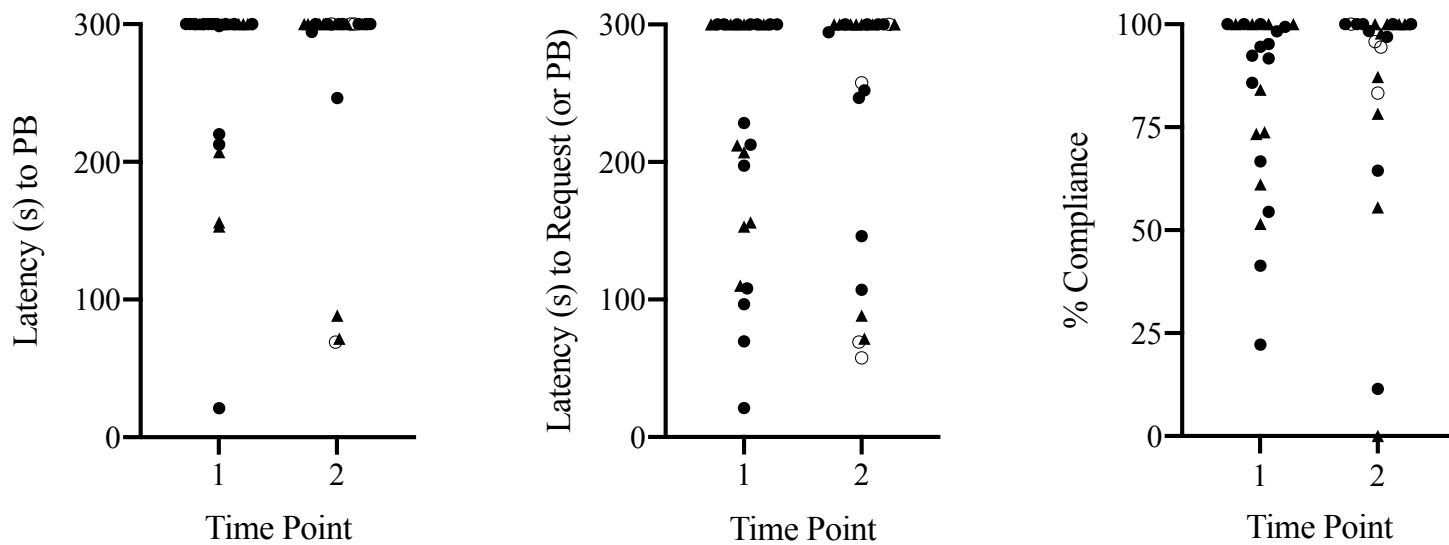
Note. PB = Problem behavior. Closed triangles represent the stimulant subgroup. Open shapes represent participants without a medication change at Time 2 (excluded from Time 2 generalizability and optimization studies).

Figure 2

Figure 2

*Projected g Coefficients Across Varying Numbers of Sessions for
Progressive Ratio (Top) and Demand (Bottom) Scores*
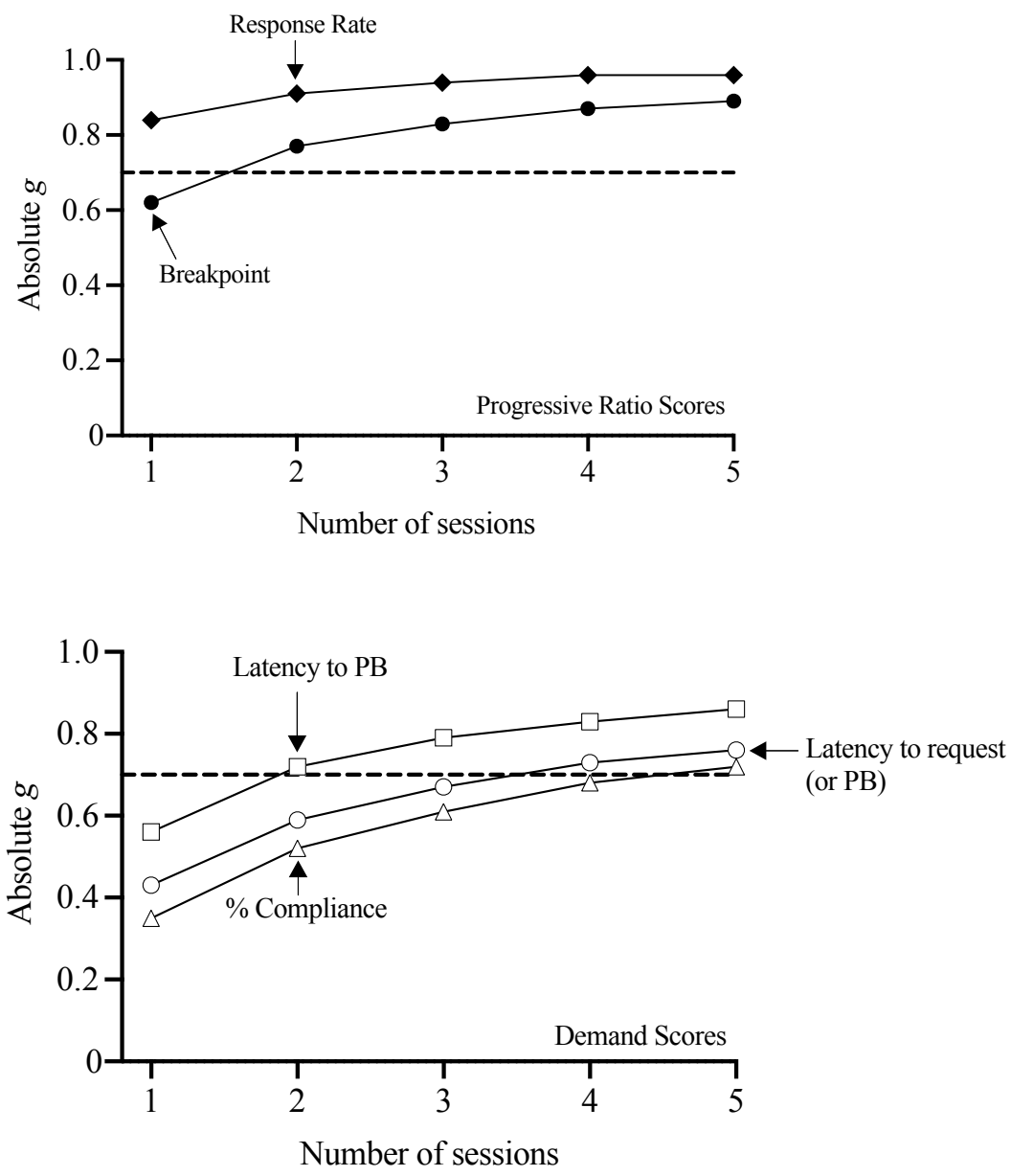
Note. Dashed line represents criterion *g* (.70). PB = Problem behavior.

Table 1

Participant Characteristics

	n	*%*
Gender		
Male	17	73.9
Female	6	26.1
Ethnicity		
White	14	60.9
Black	6	26.1
Hispanic/Latino	3	13.0
Diagnoses		
Autism	15	65.2
Attention deficit/hyperactivity disorder	13	56.5
Anxiety disorder	5	21.7
Down syndrome	3	13.0
Other genetic disorders	2	8.7
Oppositional defiance disorder	1	4.3
Mood disorder	1	4.3
Parent-Reported Topographies of Problem Behavior		
Disruption	16	69.6
Active noncompliance	15	65.2
Aggression	10	43.5
Self-injury	4	17.4
Inappropriate language	4	17.4
Elopement	3	13.0
Medication Classes (Time 1; *n* = 23)		
None	4	17.4
Stimulant	9	39.1
α_2-adrenergic agonists	6	26.1
Selective serotonin reuptake inhibitor	6	26.1
Serotonin-norepinephrine reuptake inhibitor	1	4.3
Atypical antipsychotic	1	4.3
Medication Classes (Time 2; *n* = 18)		
α_2-adrenergic agonists	11	61.1
Stimulant	9	50.0
Selective serotonin reuptake inhibitor	8	44.4
Serotonin-norepinephrine reuptake inhibitor	1	5.6
Selective norepinephrine reuptake inhibitor	1	5.6
Atypical antipsychotic	1	5.6

Table 2

Descriptive Statistics for Behavioral Assessment Scores

Behavioral Assessment Score	*N*	Mean	*SD*	Min	Max
All participants at Time 1					
Progressive Ratio					
Breakpoint	23	21.13	14.54	1	40
Response Rate	23	15.38	9.95	1.30	33.20
Demand					
Latency to PB	23	263.84	71.60	21	300
Latency to Request (or PB)	23	220.51	90.93	21	300
Percentage Compliance	23	82.00	22.47	22.22	100
Medication change participants at Time 1					
Progressive Ratio					
Breakpoint	18	21.33	14.83	1	40
Response Rate	18	14.83	9.97	1.30	32.17
Demand					
Latency to PB	18	253.80	78.38	21	300
Latency to Request (or PB)	18	213.87	93.89	21	300
Percentage Compliance	18	79.71	24.48	22.22	100
All participants at Time 2					
Progressive Ratio					
Breakpoint	21	22.97	14.97	1	40
Response Rate	21	18.52	10.96	1.53	37.17
Demand					
Latency to PB	22	266.80	78.37	69.00	300
Latency to Request (or PB)	22	235.92	94.27	57.67	300
Percentage Compliance	22	84.72	28.38	0	100
Medication change participants at Time 2					
Progressive Ratio					
Breakpoint	17	21.45	15.18	1	40
Response Rate	17	17.85	11.15	1.53	34.73
Demand					
Latency to PB	18	272.26	71.13	71.67	300
Latency to Request (or PB)	18	250.33	83.55	71.67	300
Percentage Compliance	18	82.79	31.05	0	100

Note. PB = problem behavior.

Table 3

Results of Generalizability Studies for Total Sample at Time 1 (23 Participants x 3 Sessions x 2

Observers)

Score	Source	df	MS	% Variance
Progressive Ratio				
Breakpoint	P	22	1273.44	62.3
(g = .83)	S	2	60.03	0.0
	O	1	0.72	0.0
	P x S	44	212.15	36.9
	P x O	22	2.24	0.0
	S x O	2	0.72	0.0
	P x S x O	44	2.24	0.8
Response Rate	P	22	595.08	84.3
(g = .94)	S	2	13.81	0.0
	O	1	0.38	0.0
	P x S	44	34.73	15.6
	P x O	22	0.07	0.0
	S x O	2	0.11	0.0
	P x S x O	44	0.08	0.1
Demand				
Latency to PB	P	22	667644.77	55.7
(g = .80)	S	2	30071.48	2.7
	O	1	256.12	0.0
	P x S	44	256419.19	37.9
	P x O	22	5897.55	0.0
	S x O	2	529.80	0.0
	P x S x O	44	11767.54	3.6
Latency to Request (or PB)	P	22	42628.11	42.5
(g = .70)	S	2	16431.80	0.5
	O	1	647.83	0.0
	P x S	44	11947.11	47.3
	P x O	22	1532.18	2.0
	S x O	2	2434.44	0.6
	P x S x O	44	831.88	7.1
% Compliance	P	22	2970.94	34.9
(g = .62)	S	2	3490.89	5.9
	O	1	103.76	0.1
	P x S	44	1040.79	54.7
	P x O	22	45.57	0.3
	S x O	2	11.25	0.0
	P x S x O	44	36.93	4.0

Note. P = Person; S = Session; O = Observer; PB = problem behavior.

Table 4

Results of Generalizability Studies at Time 1 vs Time 2 for Participants with Medication Change

(17/18 Participants x 3 Sessions x 2 Observers)

Score	Source	df	% Variance Accounted For	
			Time 1	Time 2
Progressive Ratio				
Breakpoint	P	16	60.9	77.2
	S	2	0.0	1.5
	O	1	0.0	0.0
	P x S	32	38.7	20.9
	P x O	16	0.0	0.0
	S x O	2	0.0	0.0
	P x S x O	32	0.3	0.4
Absolute *g*			.82	.91
Response Rate	P	16	83.0	87.5
	S	2	0.0	0.0
	O	1	0.0	0.0
	P x S	32	16.9	12.5
	P x O	16	0.0	0.0
	S x O	2	0.0	0.0
	P x S x O	32	0.1	0.0
Absolute *g*			.94	.95
Demand				
Latency to PB	P	17	56.9	82.7
	S	2	5.5	0.0
	O	1	0.0	0.0
	P x S	34	37.6	17.3
	P x O	17	0.0	0.0
	S x O	2	0.0	0.0
	P x S x O	34	0.0	0.0
Absolute *g*			.80	.94
Latency to Request (or PB)	P	17	46.6	53.8
	S	2	4.6	0.0
	O	1	0.0	2.2
	P x S	34	40.6	12.8
	P x O	17	2.0	16.2
	S x O	2	0.5	0.0
	P x S x O	34	5.6	15.0
Absolute *g*			.73	.77
% Compliance	P	17	34.4	90.6
	S	2	10.2	0.0
	O	1	0.0	0.0
	P x S	34	52.5	6.9

P x O	17	0.1	0.0
S x O	2	0.0	0.1
P x S x O	34	2.7	2.4
Absolute *g*		.62	.97

Note. P = Person; S = Session; O = Observer; PB = problem behavior.

Table 5

Pearson Product-Moment Correlations Between Behavioral Assessment Scores and ABC-2

Subscale Scores at Time 1 (n = 23) and Time 2 (n = 21 [Progressive Ratio]; n = 22 [Demand])

	ABC-2 Subscales				
	Irritability	Hyperactivity/ Noncompliance	Stereotypy	Inappropriate Speech	Social Withdrawal
Breakpoint					
Time 1	-.447*	-.564**	-.385	-.218	.060
Time 2	-.112	-.458*	-.197	-.485*	.046
Response Rate					
Time 1	-.442*	-.693**	-.467*	-.165	.165
Time 2	-.123	-.656**	-.317	-.280	.220
Latency to PB					
Time 1	-.390	-.511**	-.408	.089	.079
Time 2	-.020	-.305	-.406	-.166	.027
Latency to request (or PB)					
Time 1	-.236	-.233	-.099	.052	.196
Time 2	.150	-.142	-.134	-.380	.035
Percentage compliance					
Time 1	-.605**	-.705**	-.669**	.028	-.166
Time 2	-.477*	-.610**	-.674**	-.371	-.112

Note. PB = Problem behavior. Asterisks denote statistical significance; $p < .05$ (*) and $p < .01$ (**).