

COMPARISON OF PROBE PROCEDURES IN THE ASSESSMENT OF CHAINED TASKS

by

JENNIFER LENZ ALEXANDER

(Under the Direction of Kevin M. Ayres)

ABSTRACT

This study used an adapted alternating treatments design to compare the effects of repeated exposure of assessment procedures for chained tasks. Specifically, the researcher compared single opportunity probe (SOP), multiple opportunity probe (MOP), and, a preliminary procedure, the natural opportunity probe (NOP). The effects were first evaluated with 12 college student participants (CSP) and then replicated with 12 secondary student participants (SSP) with developmental disabilities. For the CSP ascending data were evident with MOP, zero-celerating for SOP, and variations in responding for NOP. The SSP generally responded with 0% correct across all probe procedures, with some responding in MOP and minimal responding in NOP. Implications of these findings suggest that both MOP and SOP present with testing threats and researchers should perhaps abandon these procedures for alternative choices. If MOP are to be used it is suggested that a minimum of five data collections occur prior to intervention. If SOP are to be used it is recommended that conclusions about the potency of the intervention are interpreted conservatively.

INDEX WORDS: Probe procedures, Chained tasks, Single case design, Measurement, Developmental disabilities, Special education

COMPARISON OF PROBE PROCEDURES IN THE ASSESSMENT OF CHAINED TASKS

by

JENNIFER LENZ ALEXANDER

B.S., The University of Georgia, 2006

M.Ed., The University of Georgia, 2010

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial
Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2014

© 2014

Jennifer Lenz Alexander

All Rights Reserved

COMPARISON OF PROBE PROCEDURES IN THE ASSESSMENT OF CHAINED TASKS

by

JENNIFER LENZ ALEXANDER

Major Professor: Kevin M. Ayres

Committee: Scott Ardoin
Anne Marcotte
Kristin Sayeski

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
August 2014

DEDICATION

I dedicate this to the two loves of my life, Brad and Grayson. Thank you Brad for pushing me and making sacrifices. I could have never done this without you and your continued support. Thank you Grayson for giving me a reason to keep going and a reason to smile everyday. I love you both to the moon and back.

ACKNOWLEDGEMENTS

There are a few people that need acknowledgement for their encouragement and guidance in helping me to complete my degree and dissertation. First, I would like to thank Dr. Kevin Ayres for instrumental help throughout the study. I appreciate your feedback from the beginning planning stages to the end in your assistance with final edits. Thank you also for the many invaluable experiences you have provided that have helped me grow significantly as a behavior analyst.

Thank you to Dr. Anne Marcotte and Dr. Kristin Sayeski for support in studying this topic. You both gave individualized feedback that undoubtedly increased the quality of the study and interpretation of results. Thank you to Dr. Jennifer Ledford for helping to advise me throughout the planning and writing of this study. Your feedback on this and other manuscripts throughout my degree program have greatly shaped my research design and writing skills. I also cannot go without acknowledging the other iSkills research team members, Sally Shepley and Katie Smith. Thank you for helping me to collect data and your collaboration in creating and designing the procedures for this study.

I would also like to thank my participants who took time to contribute to this study. To the college students, I hope that the results of this study may somehow influence your work with individuals with disabilities. To the secondary student participants, I hope that the work here somehow influences your future acquisition of skills that help you to exceed all expectations. I would also like to thank all of the individuals I have had the pleasure of working with while

pursuing this degree. To the individuals with disabilities, their teachers and parents, you continue to teach me more and more about life and I am blessed to have been apart of yours.

Lastly, I would like to thank my friends and family for being my cheerleaders and support system. Thank you to my sister, Becca Lenz, you are truly inspiring and one of the best teachers I know. Thank you to my mom, Carol Hunt, you have always believed in me. Thank you to my dad, Bob Lenz, you instilled me with confidence. And to my best friend Katie, I could not have done it without you. Thank you for your support, guidance, and words of wisdom and always making me laugh throughout this crazy journey. Here's to finally finishing and the next chapter of our life!

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	v
LIST OF TABLES	ix
LIST OF FIGURES	x
CHAPTER	
1 INTRODUCTION AND LITERATURE REVIEW	1
Internal Validity	2
Measurement Threats and Single Case Design.....	3
Probe Procedures	6
Current Study	13
2 METHODS	14
Participants.....	14
Settings and Arrangements	17
Materials	18
Response Definitions and Recording Procedures	19
Experimental Design.....	20
Interobserver Agreement	21
Procedural Reliability	21
Procedures.....	22
Social Validity	23

3	RESULTS	30
	Interobserver Agreement	30
	Procedural Reliability	30
	Probe Procedure Comparisons for College Student Participants.....	31
	Probe Procedure Comparisons for Secondary Student Participants	34
	Social Validity	36
4	DISCUSSION	52
	College Student Participants	52
	Secondary Student Participants.....	54
	Implications for Research	55
	Limitations	57
	Conclusion	59
	REFERENCES	61
	APPENDICES	67
	A RECRUITMENT SCRIPT	68
	B PARTICIPANT INFORMATION SHEET	69
	C GRID FOR PROBE SESSIONS.....	71
	D RUZZER DATA SHEET	72
	E LIFTON DATA SHEET.....	73
	F GALTEE DATA SHEET	74
	G PROCEDURAL FIDELITY DATA SHEET	75
	H SCREENSHOTS OF SOCIAL VALIDITY QUESTIONNAIRE ON GOOGLE FORMS.....	76

LIST OF TABLES

	Page
Table 1: College Student Participants' Information	24
Table 2: Secondary Student Participants' Information	25
Table 3: Description of Blocks in Each Set	26
Table 4: Description of Possible Errors	27
Table 5: Grouping Assignments	28
Table 6: Description of Probe Procedures	29
Table 7: Interobserver Agreement Data for College Student Participants	39
Table 8: Procedural Reliability Data for College Student Participants	40
Table 9: Procedural Reliability Data for Secondary Student Participants	41
Table 10: College Student Participants' Reported Steps Versus Observed Steps Correct	42

LIST OF FIGURES

	Page
Figure 1: CSP data during SOP trials arranged in columns by task	43
Figure 2: CSP data during MOP trials arranged in columns by task	44
Figure 3: CSP data during NOP trials arranged in columns by task.....	45
Figure 4: CSP data during SOP (open circles), MOP (open triangles), and NOP (open squares) trials arranged by participant groupings	46
Figure 5: Mean data by session for CSP (top graph) and SSP (bottom graph)	47
Figure 6: SSP data during SOP trials arranged in columns by task.....	48
Figure 7: SSP data during MOP trials arranged in columns by task	49
Figure 8: SSP data during NOP trials arranged in columns by task.....	50
Figure 9: SSP data during SOP (open circles), MOP (open triangles), and NOP (open squares) trials arranged by participant groupings	51
Figure 10: Illustration of facilitative testing threat inherent with MOP	60

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Single case design (SCD) is an experimental methodology for evaluating functional relations between dependent and independent variables. In comparison to other scientific methods in behavioral sciences (e.g., group and correlational designs), SCD relies on *baseline logic* where the individual participant serves as his or her own control (Gast & Ledford, 2014). Researchers collect data on the dependent variable in a time-series fashion across control (e.g., baseline condition) and comparison (e.g., intervention condition) conditions. Studies demonstrate experimental control by the systematic manipulation of the independent variable through (a) introduction and withdrawal, (b) staggered introduction, or (c) rapid manipulations between conditions (Horner et al., 2005). Numerous texts exist describing the selection of and procedures for designing and executing SCD studies (e.g., Gast & Ledford, 2014; Kazdin, 2012; Kennedy, 2005). Researchers also continue to publish literature on the use of quality indicators for evaluating the rigor of SCD studies (e.g., Horner et al., 2005; Kratochwill et al., 2010; Wendt & Miller, 2012) to aide in the identification of evidence-based practice (EBP).

The field of special education has recently devoted considerable attention towards identifying and cataloging EBP for individuals with disabilities. Not surprisingly, most of these summaries include SCD literature, given the benefits of using the methodology with individuals with disabilities. Interventions evaluated through group design mainly focus on determining effects through measures of central tendency. This process may not take into consideration a handful of individuals where treatment was ineffective because of possible learning differences

or disabilities. Additionally, when working within low-incidence populations, it may be difficult to identify a homogenous group large enough to evaluate statistical significance of intervention (Gast & Ledford, 2014). Hence SCD fits nicely within the evaluation of treatments and interventions in the field of special education.

Internal Validity

Defined within education, EBP are “practices and programs shown by high-quality research to have meaningful effects on student outcomes” (Cook & Odom, 2013, p. 136). When evaluating studies as *high quality*, the focus is on internal validity where confidence is related to making a case that the independent variable was responsible for changes in the dependent variable (Ventry & Schiavetti, 1986). Level of internal validity is reliant on the ability of the researcher to demonstrate a functional relation while controlling for confounding variables that could be responsible for changes in the dependent variable (Cooper, Heron, & Heward, 2007). To obtain high levels of experimental rigor, researchers have to both control for and monitor possible threats to internal validity. To increase believability in the results, the researchers should also report known limitations with their experimental design where internal validity was compromised. An intervention that has substantial literature of high quality supporting its effectiveness is then determined as an EBP.

In any experimental study, there are numerous threats to internal validity; some of which occur more often or are of greater concern in some methodologies than others. Within SCD there are threats related to variables influencing or resulting from the individual participant (e.g., history, maturation, attrition), measuring the dependent variable (e.g., testing, instrumentation), implementation of the independent variable (e.g., procedural infidelity), and experimental design decisions (e.g., multiple-treatment interference, data instability). For descriptions of individual

threats relative to SCD, see Gast and Ledford (2014). While all threats to internal validity of studies are important in SCD and identification of EBP, the remaining discussion focuses specifically on measurement threats.

Measurement Threats and Single Case Design

Without trustworthiness in the measurement procedures and data obtained, determining the effects of an intervention is difficult. Faulty measurement descriptions and procedures cause threats to internal, external, and content validity among others. Three threats related to the measurement of the dependent variable are discussed here, instrumentation, testing, and content validity.

Instrumentation. Instrumentation is related to the degree researchers collect data on the relevant dependent variable precisely and reliably throughout the study (Cooper et al., 2007). More specifically, threats are related to (a) accuracy (i.e., relation between value observed and value occurred), (b) reliability (i.e., level of consistency in values obtained over multiple measures), and (c) validity (i.e., relevance of what is measured compared to what is manipulated and reported; Kahng, Ingvarsson, Quigg, Seckinger, & Teichman, 2011). To increase internal validity with respect to instrumentation, researchers develop procedures to guard against and monitor errors related to accuracy, reliability, and validity.

Given that human observers collect the majority of data in SCD, errors are likely to occur. To guard against such threats related to accuracy and reliability, researchers write precise operational definitions of the dependent variable, select competent observers, and conduct explicit and ongoing training of data collectors (Cooper et al., 2007). To assess reliability, a second observer collects data on the same measures as the primary observer, and these values are compared. The amount of agreement between the two observers is then reported in manuscripts

as interobserver agreement (IOA; Ayres & Ledford, 2014). Low IOA is a threat to internal validity that informs the researcher of possible measurement problems related to instrumentation.

With instrumentation and measurement validity, researchers attempt to avoid threats to internal validity by conducting direct, continuous measurement. Direct measurement is favorable as it relies on observing the behavior as it occurs in comparison to indirect, which relies on individuals conveying events from memory (Cooper et al., 2007). Likewise, continuous measurement is advantageous as it attempts to measure all occurrences of the behavior within an observation period in contrast to discontinuous where researchers obtain estimated values from scheduled recordings within an observation period (e.g., interval recording; Johnston & Pennypacker, 2008).

Testing. Testing threats may occur when repeated measurement inhibits or facilitates the value of the dependent variable (Gast, 2014). Inhibitive effects are when the value of the behavior is suppressed or diminishes from measurement probe procedures. Researchers avoid inhibitive threats by interspersing known stimuli, reducing number of sessions (i.e., multiple probe design versus multiple baseline design), providing reinforcement for correct responding or related behaviors (e.g., sitting in seat, participating), and limiting participants' response effort or time in probe sessions (Gast, 2014). If inhibitive testing threats are not controlled for in baseline, suppression in responding can result in a more immediate effect when intervention is introduced than would occurred otherwise. Visual analysis in this scenario produces attribution of an inflated level of effect to the independent variable.

Facilitative effects are when the value of a behavior improves from repeated measurement in probe sessions. Researchers control facilitative threats by withholding reinforcement for correct responses, withholding prompting for correct responses, and

withholding feedback for incorrect responses (Gast, 2014). Uncontrolled facilitative threats can result in baselines in which the dependent variable is improving or becomes inflated. This inflation can result in a deflated interpretation of the effects of the independent variable may occur. Additionally, the occurrence of an improving baseline can prevent or delay the participant from receiving intervention, given the necessity of a stable baseline prior to intervention beginning (Cooper et al., 2007).

Decisions made about experimental procedures used to control for inhibitive and facilitative testing threats are oftentimes reliant on the experimenter's professional opinions, experiences, or suggestions found in texts. For example, Gast (2014) suggests that reinforcement should be withheld for correct responding in baseline when assessing receptive identification, in comparison to providing reinforcement for correct responses for expressive labeling. These types of suggestions are logical and helpful, but to advance the field of SCD in best practices for measurement decisions, research is needed on the types of measurement threats that can be attributed to specific characteristics (e.g., participants, independent variable, dependent variable).

Content Validity. Content validity is another threat that affects the trustworthiness of the measurement procedure to obtain the relevant value of the dependent variable. Generally speaking, content validity is the ability of the measurement system to sample the behavior of concern (Ventry & Schiavetti, 1986). In SCD it refers to the degree that the measurement procedure captures the true value of a participant's ability (Gast, 2014). Like the other threats discussed related to measurement, faulty content validity can negatively affect how results are interpreted within and across conditions (e.g., baseline, intervention). This may ultimately lead to underestimating or overestimating the potency of the intervention (Johnston & Pennypacker, 2008).

Probe Procedures

Within SCD there are some aspects of measurement that are generally accepted which compromise the validity of data and, in turn, a researcher's ability to conclude that changes in the dependent variable are the result of some change in the independent variable. This paper will focus attention specifically on concerns with data collection on chained behaviors. Cooper et al. (2007) defined chained behaviors as those that require a "specific sequence of discrete responses, each associated with a particular stimulus condition" (p. 435). Much of the applied literature on daily living and vocational instruction for individuals with disabilities focuses on chained behaviors that occur in a sequence such as leisure skills (Hammond, Whatley, Ayres, & Gast, 2010), food preparation (Godsey, Schuster, Lingo, Collins, & Kleinert, 2008), self-care (Ersoy, Tekin-Iftar, Kircaali-Iftar, 2009), safety skills (Wright & Wolery, 2014), or completing office tasks (e.g., Smith et al., 2013). Researchers may face challenges in measuring an individual's performance of a chained behavior given that it is inherently more complex than measuring a single discrete response (Noell, Call, & Ardoin, 2011).

Typical measurement procedures for chained tasks, like those described above, involve one of two strategies: single opportunity probes (SOP) or multiple opportunity probes (MOP) (Cooper et al., 2007; Snell & Brown, 2000). With a SOP approach, the researcher presents the participant an opportunity to perform the first step of the task. The session continues until the participant engages in an error or completes all steps correctly. If the participant makes an error, the researcher scores the error and all subsequent steps as incorrect despite the participant not having an opportunity to perform all steps. In a MOP, the participant has an opportunity to attempt all steps in the task analysis. Upon an error, the researcher arranges the environment out of view of the participant and then gives an opportunity to complete the subsequent step. The

session ends when either the participant or the researcher performs the last step in the chain. The individual can therefore demonstrate his or her ability on each discrete response despite not necessarily performing all steps correctly in sequence (Snell & Brown, 2000). With both SOP and MOP, certain threats to validity are possible and are typically documented anecdotally in the literature.

Single opportunity probe. As previously mentioned, in a SOP the researcher presents a task to a participant and as soon as the participant makes an error, the probe ends. Possible rationales for using SOP in research include collection of data representative of ability in the natural environment, time and cost efficiency, and guarding against facilitative and inhibitive testing threats to internal validity. First, if the participant is unable to independently perform the first step of a task in the natural environment, the participant will not have the opportunity to perform the second or subsequent steps. Collins, Stinson, and Land (1993) used a SOP instead of a MOP to assess participants' ability to cross the street and use a pay phone because of safety concerns associated with making errors on individual steps of these tasks. For many tasks, the first step is critical. If a participant cannot open the twist tie on the bag of bread, the participant will not be able to take out two pieces or finish making a sandwich. Second, the SOP procedure is practical to perform in terms of time and cost (Moon, Inge, Wehman, Brooke, & Barcus, 1990; Schuster, Gast, Wolery, & Guiltinam, 1988). Task analyses can include many steps that each take time to perform, but SOP last only as long as it takes to initiate and complete correct steps as specified in the measurement procedures. When SOP are conducted during the treatment condition, more time for implementation of treatment is available in comparison to MOP (Snell & Brown, 2000). Similarly, SOP result in more cost efficiency, because when a step is performed incorrectly, materials for subsequent steps are not used (Godsey et al., 2008). Lastly,

SOP guard against the specific facilitative and inhibitive testing threats, which may occur in MOP (Farlow, Loyd, & Snell, 1988; Schuster et al., 1988). In MOP, a facilitative testing effect can occur from the participant learning to perform the steps and responding correctly in subsequent sessions (see Hammond, 2011). Inhibitive testing effects can occur by the participant ceasing to respond after repeated and unnatural pausing of the session and extended exposure to lengthy sessions without the availability of reinforcement for correct responses.

Although SOP can guard against certain facilitative and inhibitive threats to internal validity, problems from the testing procedure can still occur. From an operant standpoint, consequence procedures can affect the participant's responding through punishment, extinction, and reinforcement. The researcher stopping the participant upon the first incorrect response (i.e., error) could punish attempting to complete steps resulting in non-responding in subsequent sessions. Additionally, if an opportunity for reinforcement of correct responses is unavailable during probe sessions, extinction may occur resulting in future non-responding (Farlow, Loyd, & Snell, 1987). Lastly, if the task demand or inability to perform the task is considered aversive to the participant, inappropriate behaviors or incorrect responses may be negatively reinforced when the session is stopped.

Beyond these threats to internal validity and consequent considerations, the analysis of graphed data from SOP may mislead interpretation. Line graphs could make it appear as if a participant can do few if any steps of a task analysis because sessions end when a participant makes one error. Suppose a researcher is teaching a participant to use a DVD player and the task requires five steps to play the DVD. The participant makes an error on the first step (e.g., press the power button) and the assessment ends. The researcher reports for that session, the participant responded correctly for 0% of the steps. In reality, the participant might have known

how to perform one or more steps (e.g., place disc on open tray) beyond the first. If the researcher begins instruction with the participant knowing how to perform all steps except the first, the possibility exists that the participant will show skill mastery after instruction on the first step. The researcher concludes that the intervention was potent because in a single intervention session the participant went from 0% to 100% accuracy. This illustrates how SOP is an issue with content validity because by not appropriately sampling baseline performance, baseline performance is suppressed, leading to an exaggerated immediacy of effect between conditions.

Multiple opportunity probe. Multiple opportunity probes address many of the weaknesses involved in using SOP because individuals have an opportunity to complete each step in the task analysis. Allowing participants to attempt each step results in a more accurate description of baseline performance and additional opportunities for reinforcement. In the context of chained behavior, the environmental arrangement following the completion of one step functions as the discriminative stimulus (S^D) for initiation of the following step. Therefore, when the experimenter arranges the environment (i.e., completes step correctly) it should *signal* the availability of reinforcement (e.g., completion of step) for responding. Completion of that step and subsequent presentations of the S^D for the next step allows the participant an opportunity to demonstrate what steps the participant can correctly perform. These opportunities to perform each step in MOP leads to a more complete depiction of baseline levels of responding. This improves the validity of the data because a researcher is able to document all of a participant's ability relative to the task analysis (Gast, 2014). Additionally, if reinforcement is being provided for correct responding in probe sessions, a MOP provides an occasion to receive reinforcement on each step in the task analysis, potentially resulting in improved task engagement and avoidance of inhibitive testing effects.

Multiple opportunity probes address many concerns with using SOP, but procedural concerns and testing threats exist. Multiple opportunity probes can be time consuming and costly. Time spent conducting MOP, as compared to SOP, reduces time allotted for implementing treatment. Schuster et al. (1988) used a combination of SOP and MOP to evaluate baseline performance for making a sandwich. They found that the MOP procedures quickly depleted materials, accounting for much higher instructional costs. Additionally, MOP procedures are often unnatural by interrupting the flow of work and asking the participants to turn around or close their eyes (Snell & Brown, 2000). The unnatural probe procedures can evoke and reinforce odd behaviors displayed by participants. For example, Schuster et al. (1988) found participants exhibiting behaviors such as repeating errors or opening and closing cabinets.

Researchers have also noted the risk of facilitative (Farlow et al., 1988) and inhibitive testing effects (Schuster et al., 1988) from using MOP. Researchers report that participant behavior improved after repeated exposure to MOP (Hammond, 2011; Smith, Ayres, Alexander, & Mataras, 2013). While improving behavior is typically the goal of applied research, SCD research employs strict guidelines for stable baseline responding prior to implementation of intervention. Acceleration of baseline data in a therapeutic direction compromises a researchers ability to evaluate a functional relation and calls into question the need for intervention at all (Cooper et al., 2007). However, while participants' behavior may improve with repeated testing, they are not receiving implementation of possible research-based and effective instructional practices while waiting for performance to stabilize in baseline. Further, if all that is required is repeated testing, then this raises question about the value of any instructional variable the researcher intended to evaluate.

Schuster et al. (1988) suggested attrition from lengthy sessions used with MOP. Specifically, the authors stated that participants “expressed discontent” and “gave up” because of repeated testing with MOP procedure (p. 177).

Even though the MOP procedure solves some problems presented by SOP, MOP can be inefficient, have both facilitative and inhibitive testing effects, can result in attrition, and teach new irrelevant behaviors. Researchers need to further evaluate if MOP are really depicting the most valid presentation of participants’ pre-intervention performance, or if in many cases, they simply cause more harm. There are weaknesses in using SOP alone and MOP alone, but there are potential strengths in combining the two procedures or using alternative procedures.

Natural opportunity probe. In 1988, Schuster et al. suggested that researchers evaluate procedures that would be most effective and efficient for assessing baseline performance. Their recommendation of further evaluation stemmed from disadvantages they found with MOP (i.e., material depletion, training costs, participant frustration, bizarre behaviors). For the first probe they used a SOP. Then, they conducted all subsequent probes with a “truer” measure by assessing each step individually. Others have abandoned the recommendation of using a MOP and end up using a SOP to reduce the likelihood of the participant “learning through baseline” (Tekin-Iftar, 2008, p. 263) or to reduce the cost associated with multiple sets of materials (Godsey et al., 2008).

To address the documented conflicts regarding existing probe procedures, researchers have proposed that natural opportunity probe (NOP) as an additional option for use in future research on chained tasks (Alexander, Ayres, Smith & Shepley, in preparation; Shepley, Smith, Ayres, & Alexander, in preparation; Smith, Shepley, Ayres, Alexander, & Davis, in preparation). A third option gives researchers and practitioners another choice for conducting probes on

chained tasks that may address concerns with MOP and NOP. In a NOP, the participant is given a set amount of time to complete the task that equals the sum of the desired initiation and completion time of each step. For example, operating a DVD player may have 5 steps. With a latency of 5 s for initiation and response time of 10 s for completion for each step, a participant is given a total of 1 min 15 s to complete the task (i.e., 5 steps X 10 s = 75 s or 1 min 15 s). The first occurrence of a step adhering to the topographical definition is scored as correct. Sequence of steps is not considered unless included as part of the topographical definition (e.g., start DVD player with DVD loaded) and the probe ends when one of following occurs: (a) the participant notifies the researcher of being finished, (b) 30s elapsed without the completion of a correct step, (c) the session timer ends, or (d) the participant completes all steps correctly.

Although preliminary, a procedure like NOP has the potential to solve problems associated with both SOP and MOP procedures. First, a participant has an opportunity to complete all steps of a task like a MOP (increasing content validity) while decreasing the likelihood the participant will learn through the procedures (facilitative effect) or stop responding (inhibitive effect). Although the opportunity exists with NOP, the environment may not be arranged (i.e., S^D) for all possible steps like MOP, which could decrease the opportunities for reinforcement. While cost and time will still be an issue with NOP when compared to SOP, these concerns may be less when compared to MOP because the researcher is not completing steps and the participant is stopped after a predetermined time without a correct response (e.g., 30s). On the other hand, cost may increase if the participant engages in a behavior multiple times or uses more than the defined amount of a material (e.g., pouring an entire container of laundry detergent in the washing machine). Lastly, NOP could provide a more natural scenario where the participant has a certain amount of time to complete a task. This is similar to what someone

would encounter on the job or even at home while trying to complete a novel chained task (e.g., working a new fax machine in your office).

Current Study

Alexander, Smith, Mataras, Shepley, & Ayres (in press) conducted a meta-analysis on SCD studies measuring chained tasks through multiple probe and baseline designs. Specifically, the researchers were interested in determining if there were differences in responding between SOP and MOP (i.e., absolute level change within and across conditions and slope within conditions). Data were analyzed and no statistical differences were found between the two sets of data. The authors concluded that possible limitations to number of articles found or selection bias in published studies were factors in the findings. Additionally, Alexander et al. proposed that the effects of probe procedures should be experimentally evaluated and compared. Therefore, the current study sought to answer the following research questions: (a) What are the testing effects of SOP, MOP, and NOP on college students' responding on chained nonsense tasks when analyzing variability, level, and trend? (b) When comparing variability, level, and trend what are the differences in college students' responding between SOP, MOP, and NOP? (c) What are the testing effects of SOP, MOP, and NOP on secondary students with developmental disabilities' responding on chained nonsense tasks when analyzing variability, level, and trend? And (d) when comparing variability, level, and trend what are the differences in secondary students with developmental disabilities' responding between SOP, MOP, and NOP?

CHAPTER 2

METHODS

To answer the research questions, the study was conducted through two separate investigations. The main differences between the two experiments are a result of differences in the two groups of participants, college student participants (CSP) and secondary student participants (SSP). Recording and experimental procedures with the two groups are identical and described together. In addition to the participant and setting descriptions, the only other sections with separation in narrative are the social validity procedures (CSP only) and results of the experiment.

Participants

Two sets of 12 participants were recruited to each of the two groups (i.e., CSP and SSP). The purpose of first recruiting and conducting the study with CSP was to replicate a pilot study previously conducted with CSP in which facilitative testing effects were evident with MOP and inhibitive with SOP. The current study differed from the pilot study, mainly by strengthening internal validity. When differences were found with the CSP, then the researchers recruited a group of SSP with which to conduct the study. The purpose of replicating with SSP was to increase the generality of findings to a population, in which assessment of chained tasks is typically conducted, thus increasing both external and social validity. Information about the recruitment process, inclusion criteria, and participants are described below.

College student participants. The researchers recruited CSP from a pool of 39 college-level individuals enrolled in courses in the special education department of a major university.

The pool included 25 undergraduate level preservice special educators and 14 graduate level preservice special educators and school psychologists. The researcher read a recruitment script aloud to the groups of students or sent it to them via e-mail (see Appendix A). Interested participants then signed a consent form and delivered it to the primary researcher (i.e., by hand or electronically). A total of 21 CSP agreed to participate. The researcher recorded the order in which CSP returned consents and elected the first 12 responders for the study. Two of the original participants failed to attend the first session. Of these, one dropped out based on limited availability, and the other one did not return e-mails requesting times for rescheduling a make-up session. Therefore, two additional individuals (i.e., 13th and 14th) were recruited from the list and participated in the study.

Inclusion criteria. To be included in the study, each individual met the following criteria: (a) enrolled as an undergraduate or graduate student, (b) enrolled in a degree seeking program for special education or school psychology, (c) blind to the procedures or purpose of the study, (d) currently a preservice special educator or preservice school psychologist, and (e) without sensory impairments that would prohibit participation in the block tasks. The researcher verified inclusion criteria through a participant information sheet completed after volunteering for the study (see Appendix B).

Participant information. Table 1 contains individual information about the CSP. Ages of CSP ranged from 21-27 (mean=23) and included a male and 11 females. Five CSP were seniors in a bachelor's level program to become special educators. The graduate level CSP participants included six students seeking master's degrees in special education or school psychology and one student seeking a doctoral degree in school psychology. Participants completed two questions related to the purpose of the study: (a) "What are the two most widely used probe

procedures for assessing chained tasks (if unsure, take a guess or answer with ‘I don’t know’)?” And (b) “If you are familiar with them, have you used one, both, or neither of the probe procedures in practice or research?” Six of the participants answered the first question correctly, and out of those, five reported using one or both in the past. It should be noted that based on the students’ courses at the time of the study, all had been exposed to lectures discussing SOP and MOP procedures.

Secondary student participants. Secondary student participants were recruited from six different self-contained classrooms serving students with autism spectrum disorder (ASD) and/or intellectual disability (ID). One of the classrooms was housed at a public high school, and the other five were located at a separate public facility for middle and high school students with severe problem behavior. Both schools were located in the same school district. After securing building level permission and notifying individual teachers about the study, the researcher sent parental permissions home to 12 students who met inclusion criteria (see below). After receiving all but three permissions, the researcher sent additional permissions home to three students meeting study criteria. After receiving parental permission, assent from individual SSP was secured before proceeding with the study.

Inclusion criteria. To be included in the study, each individual met the following criteria: (a) eligible for special education services under ASD, moderate ID, or mild ID as specified in their individualized education program (IEP), (b) corrected or uncorrected vision and hearing within normal limits as specified in their IEP, (c) objective related to learning pre-vocational or vocational tasks as specified in their IEP or Transition Plan, (d) intellectual functioning within the moderate or mild range as specified by psychological report, (e) fine motor ability to stack 2 single mega blocks based on a probe conducted by the primary researcher, (f) ability to follow

directions to open and close eyes based on probe conducted by the primary researcher, and (g) ability to sit for 5 min based on teacher report.

Participant information. Table 2 contains individual information about the SSP. Secondary student participants ranged in age from 13-21 (mean=17) and included 10 males and 2 females. Four of the participants were enrolled in a public school and the other eight attended a separate public school. Grades of participants ranged from 8th-12th. Nine of the SSP had an eligibility of ASD and the other three participants were served under ID (i.e., two moderate and one mild) eligibilities. Cognitive scores ranged between 40-72 and eight of the participants had a Behavior Intervention Plan (BIP). Problem behaviors addressed in the BIPs included aggression, disruptions, yelling, elopement, and off task.

Settings and Arrangements

All sessions for CSP took place in one of five rooms located on the special education floor of the College of Education building. Each room included a table and one to three chairs. Additional materials used by other individuals occupying the space were present but minimal (e.g., dry erase markers, rolling cart with tissues, paper towels, sanitizing wipes, video cameras not in use, file cabinet). Four of the rooms also included a one-way mirror connected to an observation room, but were not used for the study. Participants sat at the table while the primary observer collected data and conducted sessions by standing to his or her right. A reliability observer stood opposite of or to the left of the participant. No other individuals were present in the clinic or observation room.

Sessions for SSP took place in different locations depending on the school. For the participants at the public high school, sessions occurred at a table in the kitchen space within the classroom. The immediate area contained cabinets and kitchen appliances and materials typically

expected. At the separate public school, sessions took place at a study carrel desk within the student's classroom, at a table in a computer lab, or at a table in a separate room used to practice home-living skills. The classrooms in the separate school included typical school materials and furniture, the computer lab included four rows of desktop computers, and the home living room included a table, chair, and bedroom furniture. Participants sat at the table or desk with the primary observer standing to his or her right or left to conduct sessions and collect data. A secondary observer occasionally stood opposite of or to the left or right of the participant to collect reliability data. Other students and staff were present in the room when sessions were conducted in the classrooms, but furniture or dividers blocked their view.

Materials

The researcher created three sets of five large blocks (i.e., Mega Bloks). Each was given a name to represent a nonsense task (i.e., *ruzzzer*, *liftton*, *galtee*). Sets were created out of a pool of 15 similar but non-identical blocks in a variety of colors (red, yellow, green, blue, white, lime, turquoise), shapes (single, long, square, slope, two) and sizes (short, tall). The researcher counterbalanced color, shape, and size across sets to reduce any variability in responding in relation to appearance of materials (see Table 3 for description of blocks for each set). Three labeled half-gallon zip-lock bags separated the blocks into sets. A black grid printed on white paper and laminated was also created to use for the sessions (see Appendix C). The grid measured 5 in. X 5 in. and was comprised of 16 (4 X 4) 1.25 in. squares, each labeled with a letter and number (i.e., A1-D4) in 36 point New Times Roman font. The researcher video recorded some sessions with an iPhone 5c mounted on a Gorilla tripod for SSP for later reliability scoring, when a secondary observer was unavailable. Pens, clipboards, timers and data sheets were also available for primary and secondary data collection (see Appendix D, E, F, and

G). The researcher also created a social validity questionnaire and disseminated it using Google Forms (see Appendix H).

Response Definitions and Recording Procedures

Data collection. One of three different trained observers collected primary data on participants' responding to steps for the three nonsense tasks (see below). Steps in the task analysis were scored correct if the participant engaged in a behavior that met the topographical definition (see data sheets in Appendix D, E, F). Steps were incorrect for SOP and MOP if a topographical, sequential, latency or duration error occurred. During SOP and NOP, observers scored steps as incorrect for all steps in which the session ended without the steps attempted or completed correctly. Table 4 displays the definitions and types of errors possible for each probe procedure. The researcher summarized and graphed percent of steps correct for each task analysis.

Task analyses. The steps to compete the three nonsense tasks were created using each set of blocks for the task analyses (see datasheets in Appendix D, E, F). Each chained task included six steps and incorporated both critical and noncritical steps (Williams & Cuvo, 1986). Critical steps are those that must be completed prior to subsequent steps; whereas, noncritical steps can occur in any order. For example, to complete the third step in a task of washing clothes in a washing machine (i.e., pour detergent into the machine), a participant must remove the cap from the detergent container (step 1, critical for step 2) and then pour the detergent into the cap (step 2). The tasks in the current study were designed to reflect functional skills in this way by creating steps that were critical to the completion of other steps. For each task, steps 1 and 2 were critical for step 3 to occur, and step 5 was critical for step 6 to occur. For example, to complete the third step in creating a ruzzer (i.e., yellow slope, with slope facing down on blue two and green long

covering C1, C2, D1, D2.), the participant must have correctly completed step 1 (i.e., green long covering D1-D4) and step 2 (i.e., blue two on C1, C2).

Each task analysis also included a manipulation of a block already placed. This was in order to mimic functional tasks in which the same material is manipulated more than once, and a step may not need to occur for the final product to be correct. For example, in a task analysis for preparing a peanut butter and jelly sandwich step 1 (i.e., open jelly jar) is necessary, but step 4 (i.e., close jelly jar) is not necessary in order to create the sandwich, but may be desirable to prevent spoiling. This is exemplified in the current study in steps 4 and 5 of creating a lifton. Step 5 (i.e., lime square turned upside down covering A3, A4, B3, B4) is necessary for the final product (and for critical for step 6 to be scored correct), but step 4 is not (i.e., lime square covering A1, A2, B1, B2). The order of critical and noncritical steps and number of steps in which placement changed was consistent across all three sets in order to maintain similar response difficulty across tasks.

Experimental Design

An adapted alternating treatments design (AATD) was used to compare the effects of SOP, MOP, and NOP on acquisition of chained nonsense tasks for all 24 participants. A typical AATD includes baseline data on all skills and then compares responding with at least two different interventions on at least two skills (Wolery, Gast, & Ledford, 2014). For the purpose of this study an AATD was used to compare the three types of baseline probe procedures by evaluating the effects they have on responding on chained tasks. Therefore, the three probe procedures (i.e., SOP, MOP, and NOP) were treated as three independent variables. Each of the three nonsense tasks and probe procedures were matched and counterbalanced into six groupings. Participants from each group were randomly assigned to one of the groupings that

remained the same throughout the study (see Table 5). This resulted in two participants from each group (i.e., CSP and SSP) assigned to the same grouping. Counterbalancing allowed the researcher to measure and compare effects of repeated exposure to each probe procedure and assist in detection of unequal difficulty of tasks. Data were collected for six sessions across all participants and probe procedures in order to allow for visual inspection and compare effects.

Interobserver Agreement

The three primary data collectors and one additional data collector served as secondary observers throughout the study. Training on data collection occurred through practice sessions and 90% agreement for all probe procedures was obtained before training was completed. Secondary observations occurred live or via video recordings (i.e., SSP only) for all probe procedures. The researcher planned for IOA data to be collected for a minimum of 20% of sessions. To calculate IOA, data from each step of the task analyses were compared and scored as agreed or disagreed. Percent agreement was calculated by dividing the number of agreements by the number of agreements plus disagreements and multiplying by 100 (Ayres & Ledford, 2014).

Procedural Reliability

To assess the researcher's adherence to the procedures of this study, a secondary observer collected procedural reliability (PR) during the same sessions in which IOA data were collected (i.e., minimum of 20%). Observed and scored researcher behaviors are available in Appendix G. The observer rated each experimenter behavior a plus for correct, a minus for incorrect, or an N/A for not applicable. To calculate PR, the number of pluses was divided by the sum of pluses and minuses and multiplied by 100 (Ayres & Ledford, 2014).

Procedures

Between one and two sessions were completed each day for one or two days a week by one of the three primary data collectors. There was at least an hour in between sessions occurring on the same day. Each session included three trials in which each of the three probe procedures was presented with its assigned task. There were six possible order combinations of probing procedures for the sessions (i.e., SOP, MOP, NOP; 123, 132, 213, 231, 312, 321) and these orders were randomly assigned and counterbalanced for each participant to minimize any sequential variability. At the beginning of the session, the researcher positioned the grid on the table in front of the participant. For each trial, the researcher placed the corresponding blocks needed for the task on the table to the left of the grid and delivered the task direction (e.g., “Create the ruzzer”). The researcher delivered praise contingent on correct responses (e.g., “Good!”) and at the end of each trial (e.g., “Awesome job working”). The session ended when the participant had been exposed to each of the three probe procedures. Specific procedures for SOP, MOP, and NOP are described below and are displayed in Table 6.

Single opportunity probe. The participant had 5 s to initiate and 5 s to complete the first step in the task analysis following the delivery of the task direction. Each correct step resulted in verbal praise and 5 s to initiate and complete the subsequent step in the task. This procedure continued until the participant engaged in an error or completed all steps correctly. When an error occurred, the researcher stopped the participant from engaging with the materials, and the SOP trial ended.

Multiple opportunity probe. Like SOP, the participant had 5 s to initiate and 5 s to complete steps in the task analysis. If the participant engaged in an error, the researcher asked the participant to close their eyes, blocked their view with a binder, and completed the step. The

participant was then asked to open his/her eyes and allowed 5 s to initiate the next step. The MOP trials ended when either the researcher or the participant completed the last step in the task analysis.

Natural opportunity probe. Natural opportunity probes began the same as SOP and MOP with an initial task direction (e.g., “Create a lifton”). Following the task direction, the participant had a total of 60 s to complete the entire task (i.e., 5 s to initiate + 5 s to complete X 6 steps). A NOP trial ended when 60 s elapsed, 30 s elapsed without a correct response, or the participant completed all steps correctly.

Social Validity

The researcher distributed a questionnaire to determine the CSP perception on the purpose, procedures and outcomes of the study. The form included 11 multiple choice and two open-ended short answer questions. The questionnaire was created in Google Forms and sent to the CSP via e-mail. Responses were automatically collected and organized into a Google Spreadsheet available to the primary researcher in Google Drive. Screenshots of the questionnaire are available in Appendix H.

Table 1

College Student Participants' Information

# ¹	Participant	Sex	Age	Highest Degree	Degree Seeking	Department	Probe Question ²	History Question ³
1	Jonas	M	23	HS	B	Special Ed	No	N/A
2	Aricia	F	26	B	M	Special Ed	Y	SOP & MOP
3	Kaylee	F	22	HS	B	Special Ed	N	N/A
4	Sandy	F	21	HS	B	Special Ed	N	N/A
5	Starla	F	23	B	M	Special Ed	Y	MOP
6	Hannah	F	21	HS	B	Special Ed	N	N/A
7	Elizabeth	F	26	B	M	Special Ed	Y	NS
8	Adele	F	24	M	D	Ed Psych	N	N/A
9	Kassandra	F	24	B	M	Special Ed	Y	MOP
10	Ellie	F	27	B	M	Ed Psych	Y	NOP
11	Toni	F	23	B	M	Special Ed	Y	None
12	Carol	F	21	HS	B	Special Ed	N	N/A

Note. ¹Refers to the grouping assignment, ²refers if the participant answered correctly when asked what the two probe procedures typically used were, ³refers to the probe procedure reported using in the past, M=male, F=female, HS=high school, B=bachelor's degree, M=master's degree, D=doctoral degree, Ed=education, N=no, Y=yes, N/A=not applicable, SOP=single opportunity probe, and MOP=multiple opportunity probe.

Table 2

Secondary Student Participants' Information

# ¹	Participant	Sex	Age	Grade	School	Eligibilities	BIP	Cognitive Score ²	Measure
1	Samual	M	17	11	Separate	ASD, SLI	Y	45	SB-5
2	Gavin	M	19	12+	Separate	ASD, SLI	Y	40 ³	PPVT-4
3	Acer	M	15	10	Public	ASD, SLI	N	44	SB-5
4	Johnny	M	17	11	Public	ASD, SLI	N	72	WISC-4
5	Sissy	F	13	8	Separate	ASD, SLI	Y	49	CTONI-2
6	Jared	M	17	11	Separate	EBD, MID, SLI	Y	50 ⁴	KABC-2
7	Juan	M	14	8	Separate	ASD, SLI	Y	53	WISC-4
8	Jacob	M	18	12	Public	ASD	N	43	WISC-4
9	Susan	F	21	12+	Separate	MOID, EBD, SLI	Y	40	WISC-3
10	David	M	21	12+	Separate	EBD, MOID, SLI	Y	50	WISC-3
11	Daniel	M	16	10	Public	ASD, SLI	N	62	WISC-4
12	Thomas	M	16	10	Separate	ASD, SLI	Y	42 ⁵	NNAT

Note. ¹Refers to the grouping assignment, ²full scale intellectual quotient (IQ) unless otherwise noted, ³verbal IQ, ⁴fluid crystalized index, ⁵ standard score, M=male, F=female, ASD=autism spectrum disorder, SLI=speech language impairment, EBD=emotional behavior disorder, MID=mild intellectual disability, MOID=moderate intellectual disability, BIP=behavior intervention plan, Y=yes, N=no, SB-5=Stanford Binet, Fifth Edition, PPVT-4=Peabody Picture Vocabulary Test, Fourth Edition, WISC-4= Wechsler Intelligence Scale for Children, Fourth Edition, CTONI-2=Comprehensive Test of Nonverbal Intelligence, Second Edition, KABC-2=Kaufman Assessment Battery for Children, Second Edition, WISC-3= Wechsler Intelligence Scale for Children, Third Edition, and NNAT=Naglieri Nonverbal Ability Test.

Table 3

Description of Blocks in Each Set

Block Name	Ruzzer	Lifton	Galtee
Slope	Green	Yellow	Blue
Long	Blue	Green	Yellow
Single	White	Red	Green
Square	Red	Lime	Turquoise
Two	Yellow	Blue	Lime

Note. Bold denotes tall blocks in comparison to regular fonts denoting short blocks.

Table 4

Description of Possible Errors

Error	Definition	SOP	MOP	NOP
Topographical	Engaging in a behavior other than that described for the step	X	X	
Sequential	Engaging in a behavior described for a step but out of order	X	X	
Latency	Elapse of 5 s without initiation following the task direction (i.e., step 1) or completion of previous step (i.e., steps 2-6)	X	X	
Duration	Elapse of 5 s after initiating without completion of the step	X	X	
No Response	Step not attempted or completed	X		X

Note. SOP=single opportunity probe, MOP=multiple opportunity probe, and NOP=natural opportunity probe.

Table 5

Grouping Assignments

Grouping Assignment	CSP	SSP	Grouping Number	Probe Procedure	Task Type
1 7	Jonas Elizabeth	Samual Juan	1	SOP	Ruzzer
				MOP	Lifton
				NOP	Galtee
2 8	Aricia Adele	Gavin Jacob	2	SOP	Ruzzer
				MOP	Galtee
				NOP	Lifton
3 9	Kaylee Kassandra	Acer Susan	3	SOP	Galtee
				MOP	Ruzzer
				NOP	Lifton
4 10	Sandy Ellie	Johnny David	4	SOP	Lifton
				MOP	Ruzzer
				NOP	Galtee
5 11	Starla Toni	Sissy Daniel	5	SOP	Lifton
				MOP	Galtee
				NOP	Ruzzer
6 12	Hannah Carol	Jared Thomas	6	SOP	Galtee
				MOP	Lifton
				NOP	Ruzzer

Note. CSP=college student participant, SSP=secondary student participant, SOP=single opportunity probe, MOP=multiple opportunity probe, and NOP=natural opportunity probe.

Table 6

Description of Probe Procedures

Single Opportunity Probe	Multiple Opportunity Probe	Natural Opportunity Probe
1. Set blocks to the left of the grid	1. Set blocks to the left of the grid	1. Set blocks to the left of the grid
2. Provide a task direction (“Create the _____”)	2. Provide a task direction (“Create the _____”)	2. Provide a task direction (“Create the _____”)
3. Allow 5 s for initiation of each step	3. Allow 5 s for initiation of each	3. Start trial timer for completion amount (i.e., initiation time of each step + completion time of each step; e.g., [5 s X 6 steps] + [5 s X 6 steps] = 60 s)
4. Allow 5 s to complete each step	4. Allow 5 s to complete each step	4. If the participant performs the step correctly, mark (+) on the data sheet, and provide praise
5. If the participant performs the step correctly, mark (+) on the data sheet, and provide praise	5. If the participant performs the step correctly, mark (+) on the data sheet, and provide praise	5. Assessment ends when <ol style="list-style-type: none"> Participant completes all steps correctly 30 s elapse without correct Timer ends
6. If the participant performs an error, mark (-) on the datasheet, and end the probe trial	6. If the participant performs an error, mark (-) on the datasheet, ask the participant to close their eyes, block view with binder, complete the step correctly, remove binder from view, ask to open eyes, and allow them to attempt the next step	6. Provide praise for participating
7. Assessment ends when <ol style="list-style-type: none"> First error occurs Participant completes all steps correctly in order 	7. Assessment ends when the last step is completed by the participant or researcher	7. Mark all steps not completed correctly as incorrect (-)
8. Provide praise for participating	8. Provide praise for participating	
9. Mark all steps not attempted, as incorrect (-)		

CHAPTER 3

RESULTS

Interobserver Agreement

Secondary data were collected for of 33% of sessions for all CSP and probe procedures and overall agreement was 99.5%. For SOP, MOP, and NOP, mean IOA was 100%, 100%, and 98.6% (range=83-100%) respectively. The two disagreements with NOP occurred on steps one and two. A lower percentage agreement with NOP when compared with the others is not surprising given the difficulty of marking correct responses that could occur out of sequence.

Table 7 displays the mean IOA for CSP by participant and probe procedure.

Secondary observers collected data for a mean of 32% of sessions across SSP and probe procedures. In 33% of sessions, secondary data were collected for majority of participants (n=10). For one participant, a second observer collected reliability data in 50% of sessions (i.e., scheduling error where the third one was not needed). For the other participant, he was removed from the study in the second session before any IOA had been collected (see results below). The overall agreement for SSP across all probe procedures was 100%.

Procedural Reliability

Procedural data were collected in the same percentage of sessions as IOA for CSP (i.e., 33%). The mean PR across CSP for SOP, MOP, and NOP was 99.4% (range=93-100%), 98.5% (range=91.5-100%), and 99.4% (range=93-100%) respectively. The overall mean PR for CSP was 99.1%. Errors during CSP sessions occurred for five participants and at least once in each probe procedure. Errors during SOP and NOP were related to not providing a praise statement

for a correct response. During MOP errors included not providing praise at the end of the trial, not allowing 5 s to initiate a step, not allowing 5 s to complete a step, and not asking the participant to close their eyes following an error. The latter was a result in the observers disagreeing in the occurrence of an error in responding. Table 8 displays the mean PR for each CSP by probe procedure.

Procedural data were collected in the same percentage of sessions as IOA for SSP (i.e., 32%). The average reliability across SSP for SOP, MOP, and NOP was 100%, 99.7% (range=97-100%), and 98.2% (range=87.5-100%) respectively. The overall mean PR for SSP was 99.3%. Errors during SSP sessions included not providing the correct task direction for MOP and not beginning a timer and not providing praise at the end for NOP. Table 9 displays the mean PR for each SSP by probe procedure.

Probe Procedure Comparisons for College Student Participants

Correct responding on the individual chained tasks for each probe procedure was measured through direct observation and evaluated through the AATD design. Percent correct for each procedure for each participant was graphed and grouped by probe procedure (i.e., Figures 1-3). The researcher visually examined data from each probe procedure for variability, level, and trend. Comparisons were then conducted by visually analyzing the data for effects of the probe procedures for individual participants and as a group (i.e., Figures 4 and 5). Interpretations of the results for CSP follow.

Single opportunity probe. Figure 1 displays the effects of repeated exposure to SOP on the CSP block assembling behavior. The graphs are arranged in columns by type of task (i.e., ruzzer, lifton, galtee). Toni was the only CSP to respond correctly during SOP sessions; she completed one step correctly on the last session. Responding in all sessions for the other eleven

CSP was at 0%. Therefore, the only trend was with Toni, where there was slight acceleration. This one data point accounted for minor variability and level change across SOP sessions.

Multiple opportunity probe. Figure 2 displays the effects of MOP on the CSP block building behavior and graphs are arranged in columns by task type. In the first session, all but one participant (i.e., Kaylee; 17%) responded with 0%, but by the second session, half of the participants increased to between 17-50% (i.e., between 1-3 steps correct; mean=16.7%). By the sixth session, half of the participants obtained 100% correct responding, with the remainder responding between 17-83% by the last session (mean=80.5%).

When visually analyzing the data, eleven of the CSP have accelerating data paths. Kaylee was the exception with a zero-celerating trend. Out of the eleven CSP with accelerating data, two did not begin responding until the last session, where Sandy completed 17% of steps correctly and Elizabeth completed 100% of steps. Variability was relatively low overall with most participants continuing to increase in correct responding. Responding decreased once for half of the CSP, but the decrease was usually by one step (67%). Carol demonstrated the greatest decrease in responding from session two to three where her responding decreased from 50% back to 0%. Her data are the most variable where within the first four sessions her responding alternated between 0% and 50% then 0% and 100%. Most decreases occurred in session three (50%), with others occurring in sessions two (17%) and five (33%).

The researcher conducted a within level analyses of the data. Eleven CSP increased their responding from the first session to the last with a mean absolute level change of 79.1% (last data point minus first data point; Gast & Spriggs, 2014). Relative level change was also calculated by subtracting the median of the first three data points from the last three data points (Gast & Spriggs, 2014). Mean relative level change across the CSP was 51.3%.

Natural opportunity probe. Figure 3 displays the effects of NOP on the CSP block building behavior and graphs are arranged in columns by task type. Overall responding ranged between 0-50%, where seven CSP responded correctly at least once in the NOP sessions, and the other five remained at 0% for all sessions. Four of the participants engaged in correct responding in one session and got one step correct (i.e., 17%). This occurred in different sessions for all four participants (i.e., 2, 3, 5, 6). Cassandra responded with 17% correct in the first two sessions then dropped down to 0% for the remainder of probes. After three sessions at 0%, Aricia responded correctly in the last three sessions at 17%, 17%, and 33% respectively. Adele had the most sessions with correct responding and the highest maximum percentage. On the first session she responded at 17%, then increased to 50% in session two and went back down to 33% in session three. After two sessions at 0%, Adele ended on session six with 50% correct.

Five CSP data were in a zero-celerating trend at 0%. Four participants have slight decelerating trends and the other three CSP have slight accelerating trends. Variability was relatively low overall. Most participants' responding remained between 0-17% correct, and for Aricia, responding continued to increase. Adele's data are the exception to this, where her data are highly variable.

Within level analyses were conducted. Aricia, Elizabeth, and Adele increased their responding from the first session to the last (i.e., 33%, 17%, 33%), and Cassandra decreased in responding (i.e., 17%). The mean absolute level change across CSP was 5.5%. Aricia was the only CSP with a relative increase in trend (i.e., 17%), with relative decreases in trend in Cassandra and Adele's data (i.e., -17%, -33%). Mean relative level change across the CSP was -2.8%.

Comparisons. Figure 4 displays individual data for all 12 participants for all probe procedures. The graphs are arranged into the groupings the CSP were assigned to by row. When looking at individual CSP data, all participants responded correctly more often in MOP than in NOP and SOP. For seven participants, responding was greater for NOP than SOP. This is also evident in Figure 5, which displays the mean data point across participants for each probe procedure.

When comparing trend across probe procedures, acceleration was more likely with MOP. When comparing NOP and SOP, acceleration more often occurred with NOP. A decelerating trend was mostly associated with NOP. All but one participant responded in a zero-celerating trend during SOP trials. Therefore, SOP is most likely to result in a zero-celerating trend, followed by NOP, then MOP.

Variability was relatively low across all probe procedures. Stable responding was almost always associated with SOP, where responding remained at 0% for majority of sessions. Some variability was observed in both NOP and MOP, but more so with MOP. When looking at changes in trend within probe procedures, absolute and relative change was the lowest with SOP (i.e., 1.4% and 0% respectively). NOP resulted in a mean absolute level increase of 5.5%, but a decrease of 2.8% when calculating relative change. MOP resulted in the greatest mean increase both when calculating both absolute (79.1%) and relative (51.3%) level changes.

Probe Procedure Comparisons for Secondary Student Participants

Correct responding on the individual chained tasks for SSP was graphed and analyzed in the same manner as CSP. The researcher visually examined the data from each probe procedure (i.e., Figures 6-8). Comparisons were then conducted by visually analyzing the data from the three probe procedures for individual participants and as a group (i.e., Figures 5 and 9). For two

of the SSP (i.e., Samuel and Juan), one session ended when the primary researcher was struck by each of the participants. Both occurrences of aggression (i.e., hit with object in face, hit with hand on chin) occurred during a MOP. This resulted in the termination of the study for both participants and is reflected in their data. Interpretations of the results for SSP follow.

Single opportunity probe. Figure 6 displays the effects of SOP on the SSP arranged in columns by type of task (i.e., ruzzer, lifton, galtee). For all SSP, in all sessions, responding remained at 0%. Therefore there was no celeration or variability in responding within individual participants or differences between participants.

Multiple opportunity probe. Figure 7 displays the effects of MOP on the CSP block building behavior and graphs are arranged in columns by task type. Only four SSP responded correctly during at least one session, while the other eight remained at 0% throughout all six sessions. Correct responding never exceeded 17% (i.e., one step correct). Johnny only had one session with correct responding; Sissy and Susan with two, and David had the most with three.

When looking at trend, the majority of participants' data were zero-celerating at 0%. For the other four SSP, their data were slightly accelerating. Variability was low to non-existent for the SSP during MOP trials with data remaining between 0-17% for all participants. When looking at within trend analyses, only Sissy had an increase from first to last data point, resulting in a mean absolute level change for SSP of 1%. Relative level change was also calculated for all participants and when averaged together resulted in a mean increase of 3%.

Natural opportunity probe. Figure 8 displays the effects of NOP on the CSP block building behavior and is arranged in columns by task type. Correct responding only occurred once each for Jacob and Susan. This occurred in sessions two and one respectively. Overall responding was zero-celerating and stable at 0%. For the two participants who responded, their

data could be considered slightly decelerating with minimal variability. Mean absolute level change for SSP was -1% and relative level change was 0%.

Comparisons. Figure 9 displays the individual data for all 12 SSP for all probe procedures. The graphs are arranged into the groupings the SSP were assigned to by row. When looking at individual SSP data most participants remained at 0% for all probe procedures and all sessions. Participants who responded correctly did so in MOP or NOP only. Although low, responding was more likely to occur with MOP than NOP. This is also evident in Figure 5, where the mean data point across SSP for each probe procedure is displayed.

When comparing trend across probe procedures, MOP was more likely than NOP or SOP to result in an accelerating trend. A decelerating trend was most likely to occur with NOP. Zero-celerating trends occurred more often than not for all probe procedures, but were most likely for SOP, followed by NOP, then MOP. Variability was low across all probe procedures with responding remaining between 0-17%. Stable responding was always associated with SOP, where responding remained at 0% for all SSP. Overall there was little change in respect to trend within the probe procedures. Mean absolute level change was 0% for SOP, -1% for NOP, and 1% for MOP. Relative change was 0% for both SOP and NOP and 3% for MOP.

Social Validity

Eleven out of twelve CSP completed the social validity survey. The first set of questions was focused on asking the participants how they performed on the different tasks on the first and last session. Comparing CSP reported values to researcher observed values helped to validate their ability to answer questions about the different tasks. Table 10 displays comparisons between the numbers of steps the CSP reported getting correctly versus what was observed. Out of 72 possible comparisons, the participant and observer agreed 83% of the time (i.e., exact

match on the number of steps correct). Out of the disagreements, 8% were with SOP, 59% MOP, and 33% NOP. For the disagreements the participants were usually off by one (i.e., 83%; range=1-3, mean=1.25).

After reporting how they performed for each task, CSP were asked to select between two options about their behavior for most of the sessions related to engaging with materials: (a) attempted to manipulate blocks or (b) did not attempt to manipulate blocks. Out of 36 possible participant and task combinations, only seven “did not attempt” were reported. Of those, 43% were SOP, 29% were MOP, and another 29% were NOP. Participants were also asked to select which probe procedure they performed the best on and all reported the correct task. When asked why they thought they performed the best on the one they selected (i.e., MOP), participants generally reported that it was because they were shown how to complete it. Examples of answers included, “[It] was the only one with immediate feedback”, “The error correction procedure in the Galtee task seemed to be helpful”, “I was given opportunities to watch how to build [it]”, and “I did the best because of the prompting”. Toni was specific about what was occurring in regard to the probe procedure, “I was given an opportunity to see the finished product through the use of the multiple-opportunity method.”

The last few questions of the social validity questionnaire were related to their opinion of what they believed they were to do for the tasks. When asked if their perception of what they were supposed to do changed during the study, 67% of CSP responded with yes. When asked what changed, participants gave answers related to what they were supposed to do with the blocks. Most participants reported learning that they were supposed to use the grid, while others spoke specific to learning to wait to see the answer in the MOP. One participant believed that there were different behaviors she was to do with each task and stated, “I figured out most of the

Galtee [MOP] over time, but I kept doing the same thing with the Ruzzer and the Lifton. I thought the point of the Ruzzer [SOP] was to touch the long block and end the session, and the point of the Lifton [NOP] was to kind of make whatever I wanted to.”

Table 7

Interobserver Agreement Data for College Student Participants

Name	SOP	MOP	NOP	Average
Jonas	100.0%	100.0%	100.0%	100.0%
Aricia	100.0%	100.0%	100.0%	100.0%
Kaylee	100.0%	100.0%	100.0%	100.0%
Sandy	100.0%	100.0%	100.0%	100.0%
Starla	100.0%	100.0%	100.0%	100.0%
Hannah	100.0%	100.0%	100.0%	100.0%
Elizabeth	100.0%	100.0%	100.0%	100.0%
Adele	100.0%	100.0%	83.0%	94.3%
Kassandra	100.0%	100.0%	100.0%	100.0%
Ellie	100.0%	100.0%	100.0%	100.0%
Toni	100.0%	100.0%	100.0%	100.0%
Carol	100.0%	100.0%	100.0%	100.0%
Average	100.0%	100.0%	98.6%	

Note. SOP=single opportunity probe, MOP=multiple opportunity probe, and NOP=natural opportunity probe.

Table 8

Procedural Reliability Data for College Student Participants

Name	SOP	MOP	NOP	Average
Jonas	93.0%	91.5%	100.0%	94.8%
Aricia	100.0%	100.0%	100.0%	100.0%
Kaylee	100.0%	100.0%	100.0%	100.0%
Sandy	100.0%	100.0%	100.0%	100.0%
Starla	100.0%	97.0%	100.0%	99.0%
Hannah	100.0%	100.0%	100.0%	100.0%
Elizabeth	100.0%	100.0%	100.0%	100.0%
Adele	100.0%	100.0%	93.0%	97.7%
Kassandra	100.0%	100.0%	100.0%	100.0%
Ellie	100.0%	97.0%	100.0%	99.0%
Toni	100.0%	97.0%	100.0%	99.0%
Carol	100.0%	100.0%	100.0%	100.0%
Average	99.4%	98.5%	99.4%	

Note. SOP=single opportunity probe, MOP=multiple opportunity probe, and NOP=natural opportunity probe.

Table 9

Procedural Reliability Data for Secondary Student Participants

Name	SOP	MOP	NOP	Average
Samual	100.0%	100.0%	100.0%	100.0%
Gavin	100.0%	97.0%	100.0%	99.0%
Acer	100.0%	100.0%	100.0%	100.0%
Johnny	100.0%	100.0%	100.0%	100.0%
Sissy	100.0%	100.0%	87.5%	95.8%
Jared	100.0%	100.0%	100.0%	100.0%
Juan	N/A	N/A	N/A	N/A
Jacob	100.0%	100.0%	100.0%	100.0%
Susan	100.0%	100.0%	100.0%	100.0%
David	100.0%	100.0%	100.0%	100.0%
Daniel	100.0%	100.0%	100.0%	100.0%
Thomas	100.0%	100.0%	93.0%	97.7%
Average	100.0%	99.7%	98.2%	

Note. SOP=single opportunity probe, MOP=multiple opportunity probe, and NOP=natural opportunity probe.

Table 10.

College Student Participants' Reported Steps Versus Observed Steps Correct

Name		Ruzzer				Lifton				Galtee				Agg			
		First		Last		First		Last		First		Last					
		R	A	R	A	R	A	R	A	R	A	R	A				
Jonas	SOP	0	0	0	0	MOP	0	0	6	6	NOP	0	0	0	0	100%	
Aricia	SOP	0	0	0	0	NOP	0	0	0	2	MOP	1	0	5	6	50%	
Kaylee	MOP	N/A	1	N/A	1	NOP	N/A	0	N/A	0	SOP	N/A	0	N/A	0	N/A	
Sandy	MOP	0	0	1	2	SOP	0	0	0	0	NOP	0	0	0	0	83%	
Starla	NOP	0	0	0	0	SOP	0	0	0	0	MOP	0	0	6	6	100%	
Hannah	NOP	0	0	0	0	MOP	1	0	6	5	SOP	0	0	0	0	67%	
Elizabeth	SOP	0	0	0	0	MOP	0	0	6	6	NOP	0	0	0	1	83%	
Adele	SOP	0	0	0	0	NOP	0	0	0	3	MOP	0	0	5	5	100%	
Kassandra	MOP	0	0	6	5	NOP	0	1	0	0	SOP	0	0	0	0	67%	
Ellie	MOP	0	0	3	4	SOP	0	0	1	0	NOP	0	0	0	0	67%	
Toni	NOP	0	0	0	0	SOP	0	0	1	1	MOP	0	0	6	6	100%	
Carol	NOP	0	0	0	0	MOP	0	0	6	6	SOP	0	0	0	0	100%	
Agg:		100%		73%		Agg:		82%		64%		Agg:		91%		82%	

Note. Bolded numbers=disagreement, R=reported by participant, A=actual data from observation, SOP= single opportunity probe, MOP= multiple opportunity probe, NOP= natural opportunity probe, N/A=not available, and Agg= agreement.

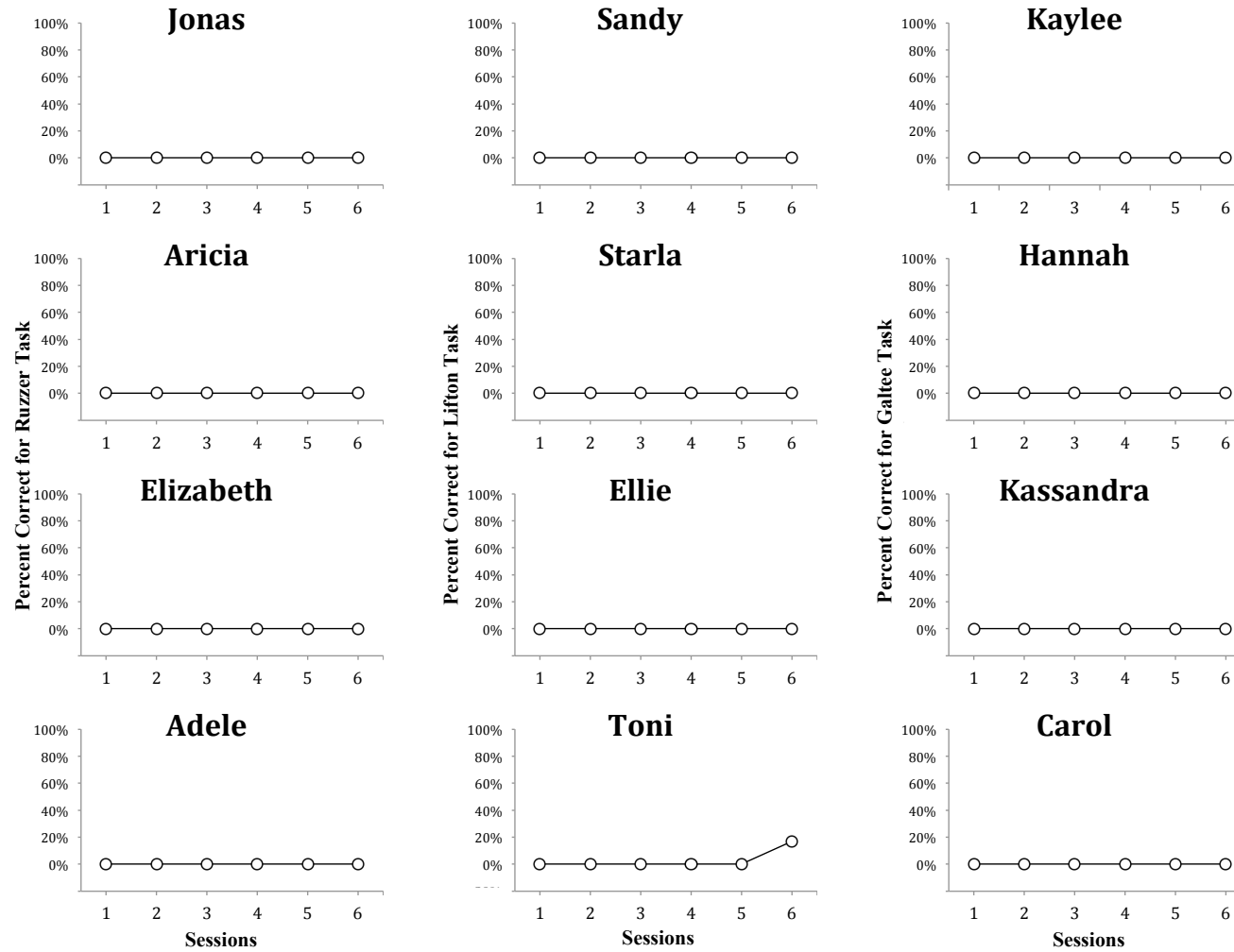


Figure 1. CSP data during SOP trials arranged in columns by task.

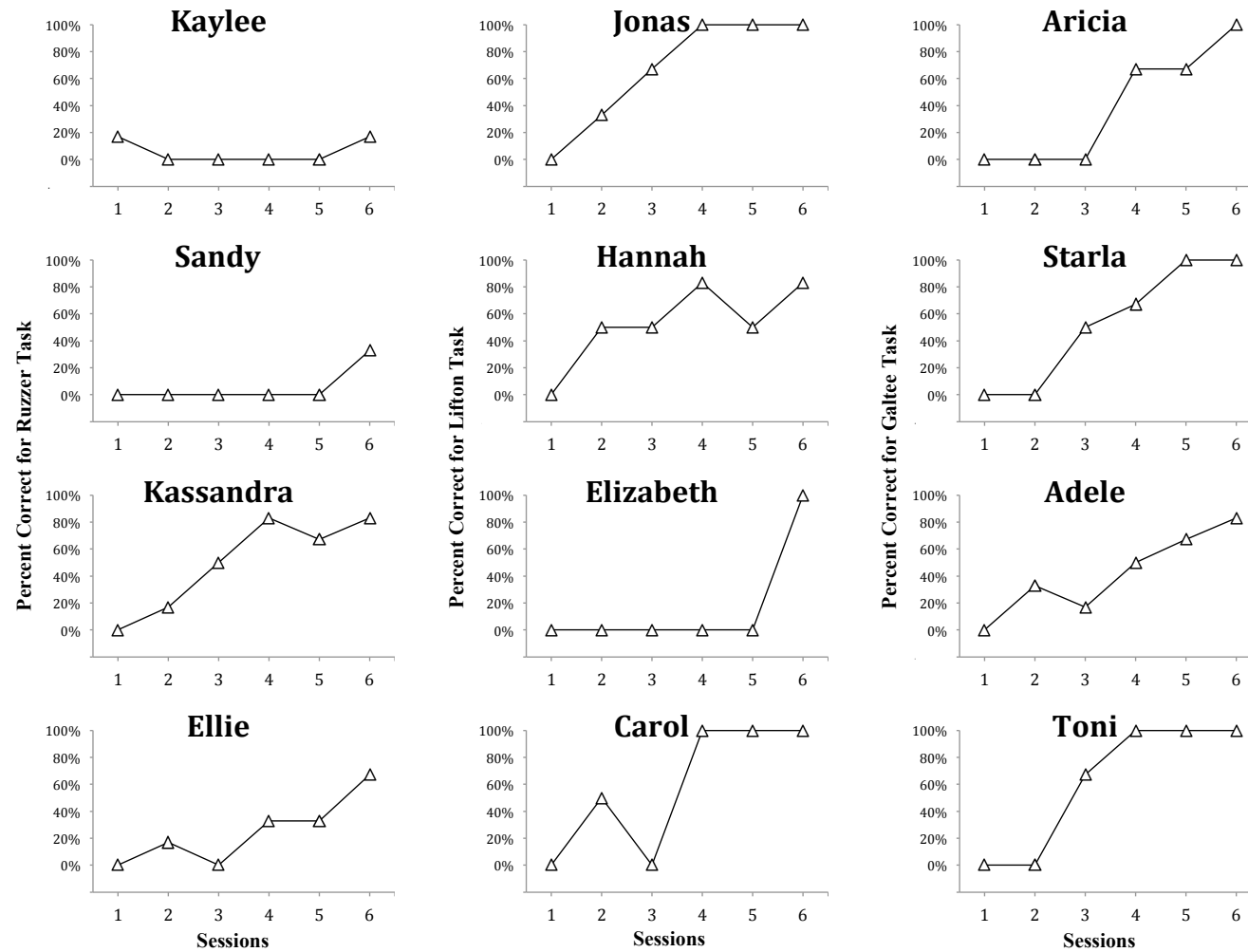


Figure 2. CSP data during MOP trials arranged in columns by task.

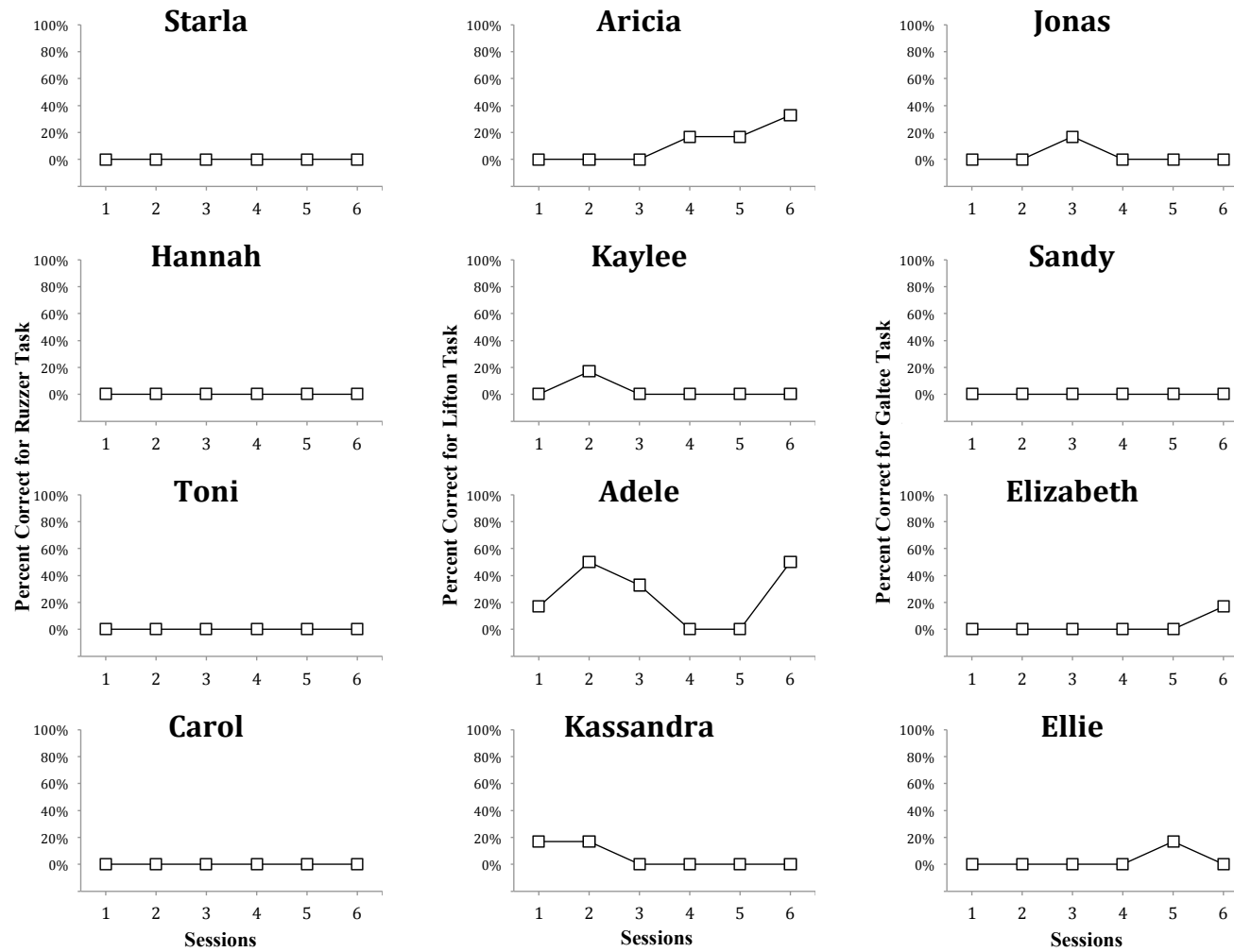


Figure 3. CSP data during NOP trials arranged in columns by task.

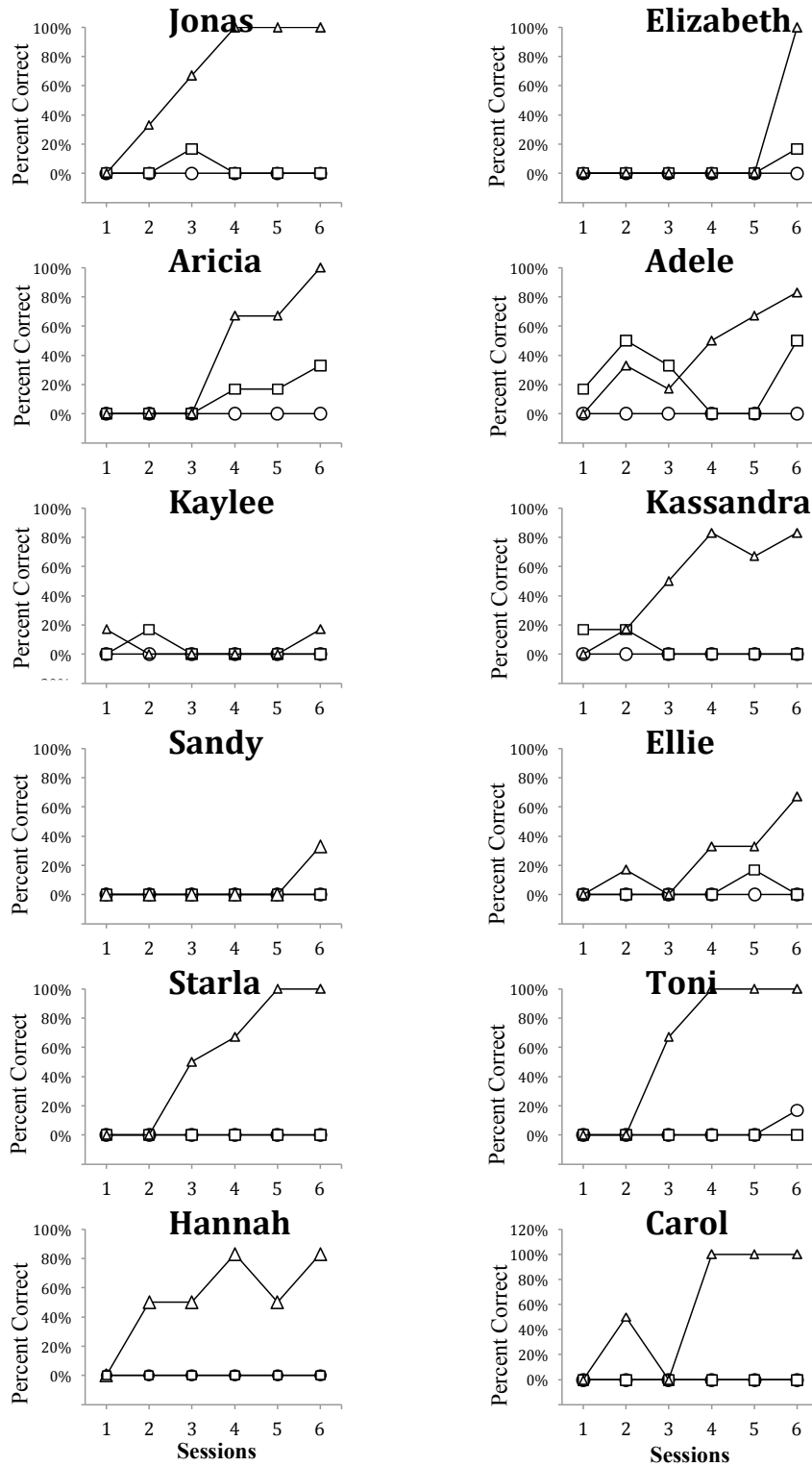


Figure 4. CSP data during SOP (open circles), MOP (open triangles), and NOP (open squares) trials arranged by participant groupings.

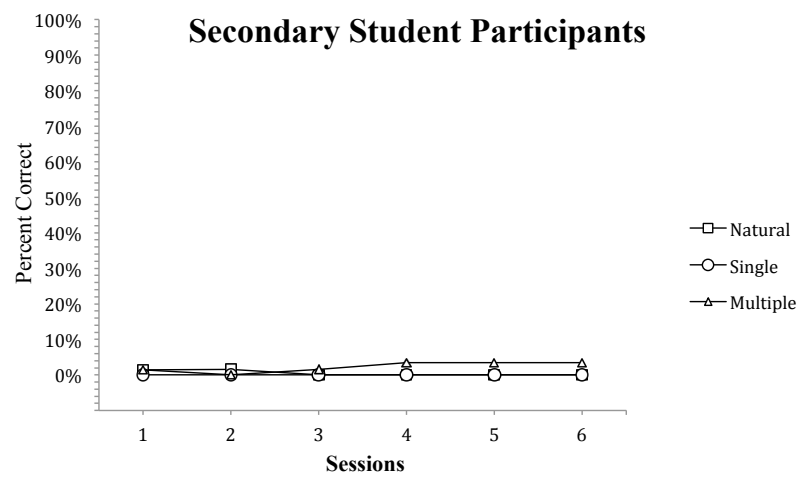
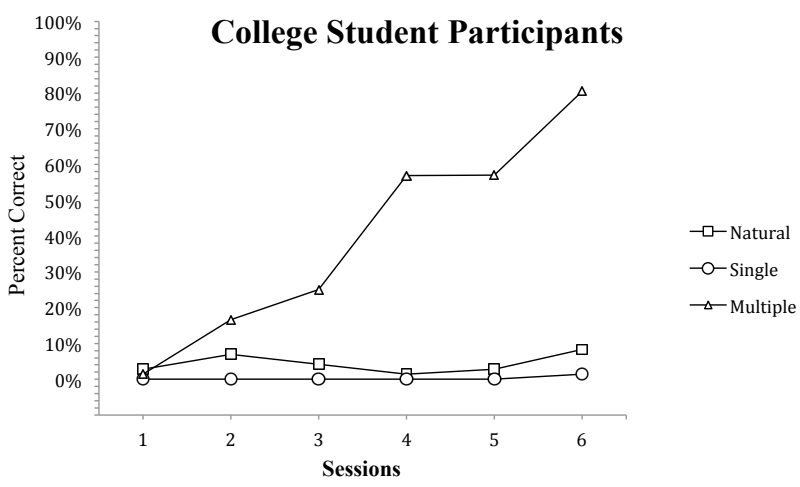


Figure 5. Mean data by session for CSP (top graph) and SSP (bottom graph). Data are displayed for SOP (open circles), MOP (open triangles), and NOP (open squares) trials.

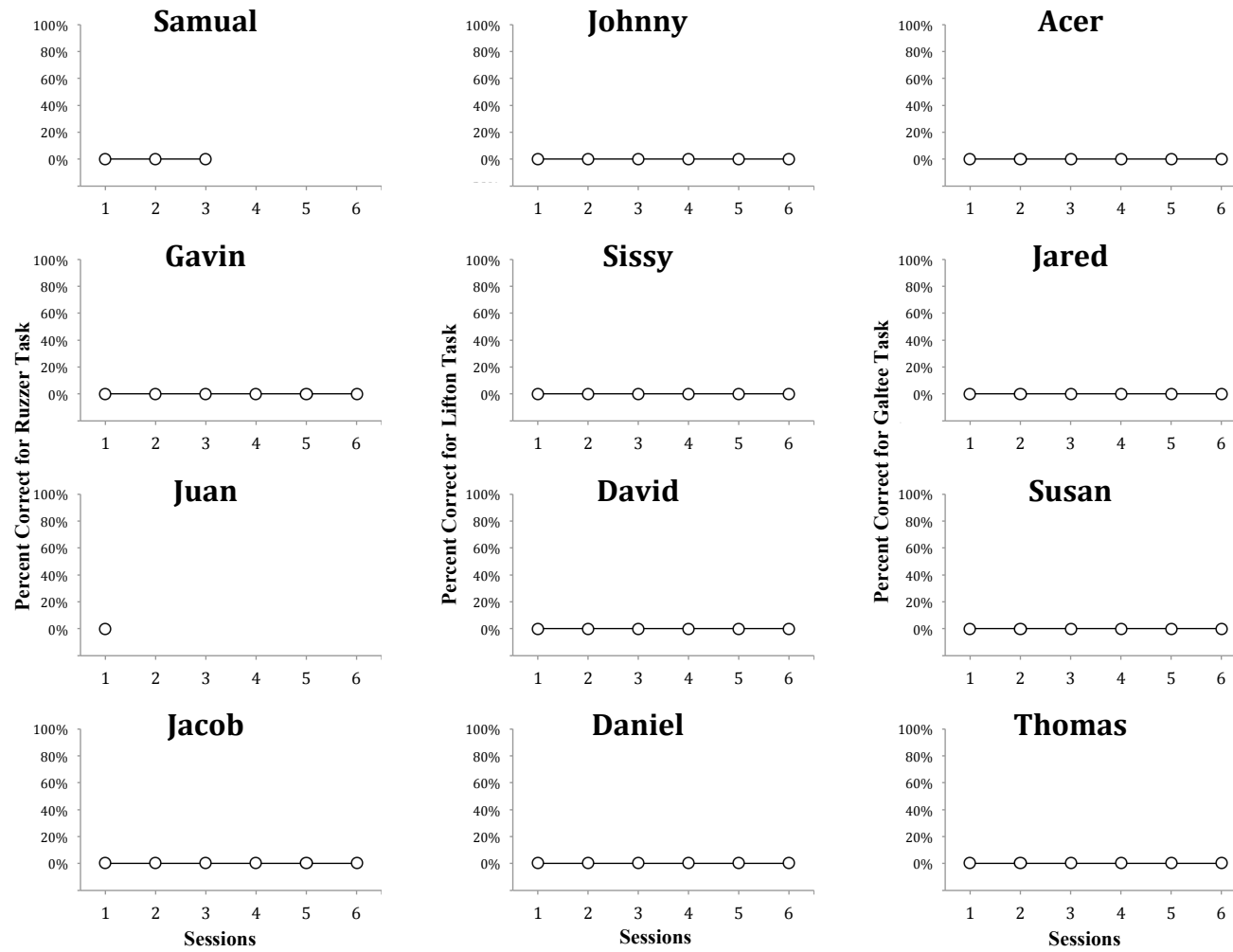


Figure 6. SSP data during SOP trials arranged in columns by task.

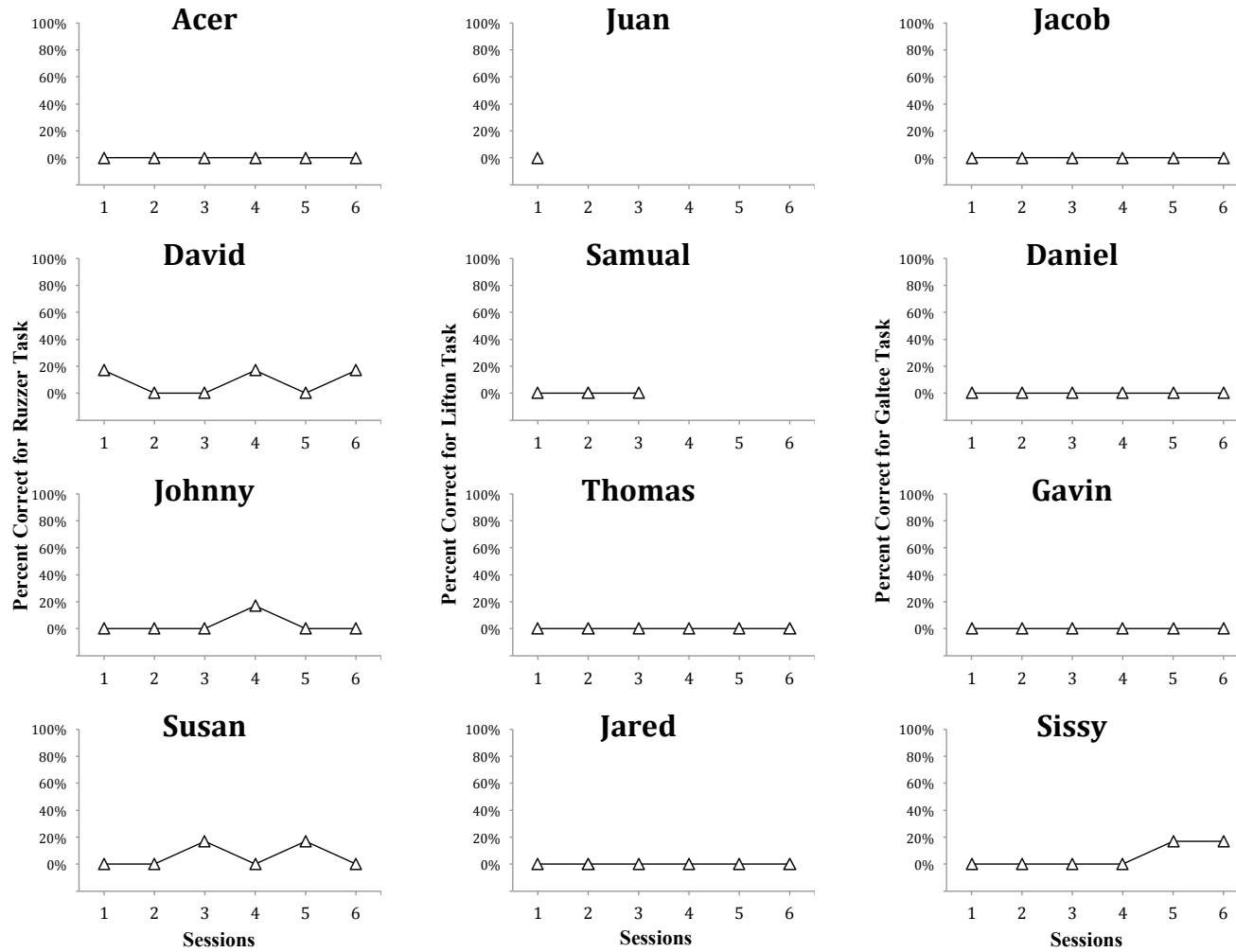


Figure 7. SSP data during MOP trials arranged in columns by task.

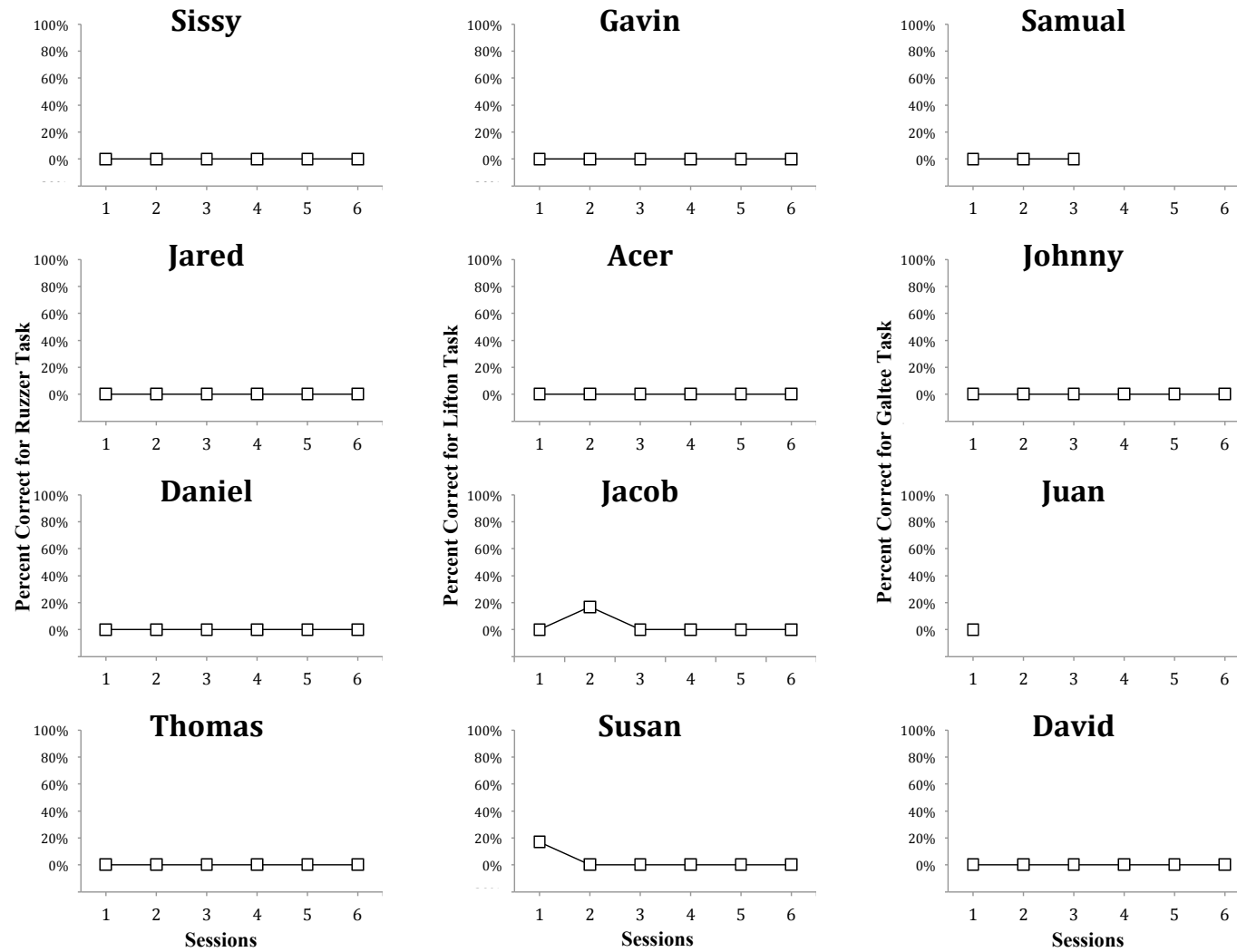


Figure 8. SSP data during NOP trials arranged in columns by task.

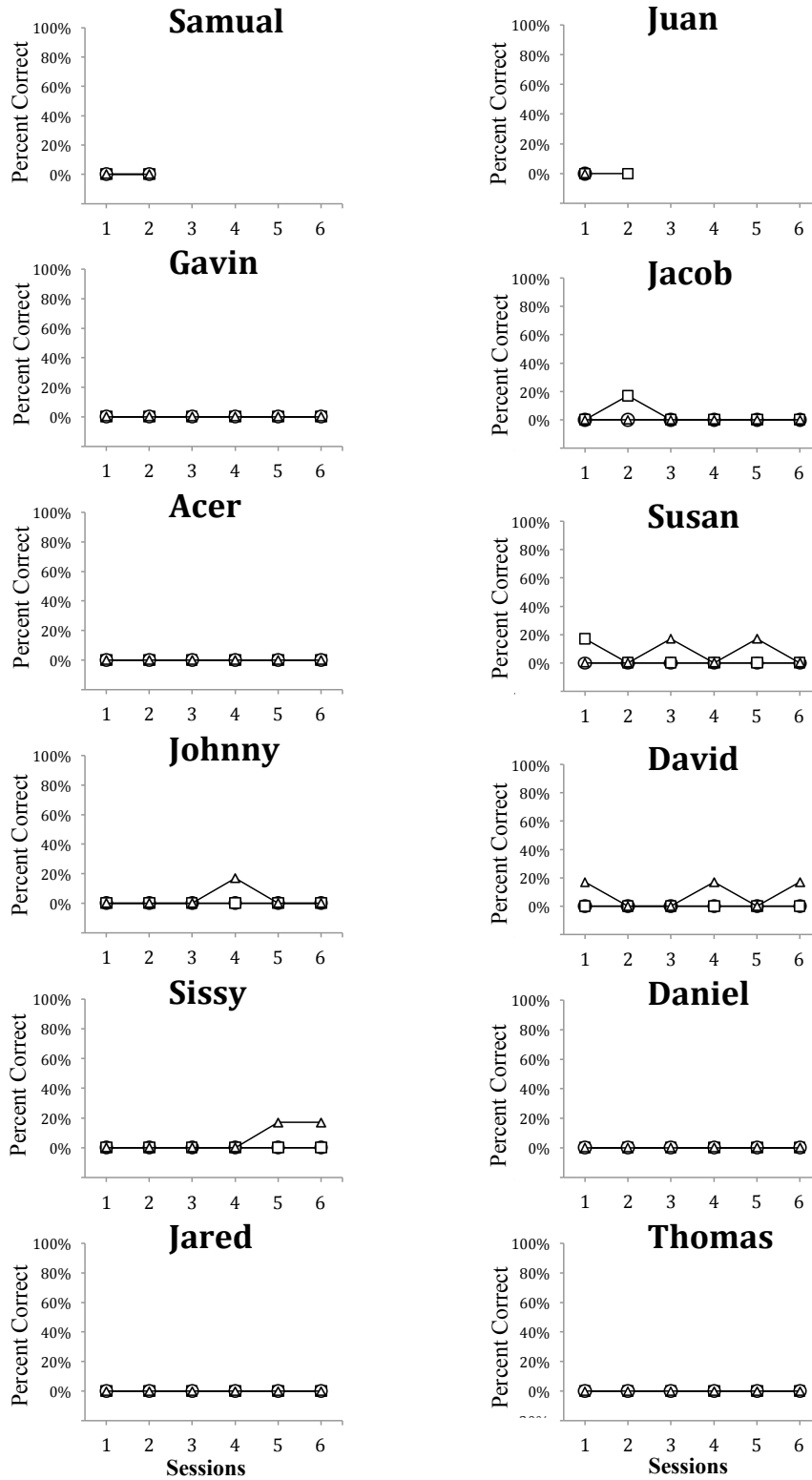


Figure 9. SSP data during SOP (open circles), MOP (open triangles), and NOP (open squares) trials arranged by participant groupings.

CHAPTER 4

DISCUSSION

College Student Participants

When examining effects of the probe procedures on nonsense tasks, there are distinct patterns in responding. First, SOP almost always resulted in zero steps scored correct. When visually analyzing the data, overall trend is zero-celerating with almost no variability or change in level. This is important because it is possible that given the opportunity to complete more than the first step (or second for Toni in one session), the participants would have had more correct responding and opportunities to receive reinforcement. This could be a possible inhibitive testing effect where suppressed responding is occurring. An additional concern related to content validity is warranted given that the SOP procedure is not sampling the true performance of the CSP. Although it is unlikely that the CSP had prior knowledge of how to build the structures created for this study, they do have block building in their repertoires. This is analogous to steps in chains that individuals may know how to complete out of sequence or within other chains. Consider two task analyses related to washing surfaces where the participant is able to perform one task at 100% (e.g., cleaning a window) and a new one is being assessed (cleaning a wooden table). Both tasks may have similar steps such as wiping the surface, but differ in the way the solution gets to the surface (e.g., Windex bottle for windows versus Pledge bottle for furniture). In this case the wiping behavior is in the participant's repertoire, but would never get assessed on the cleaning furniture task with a SOP if the participant performs the spray step (not in participant's repertoire) incorrectly.

All CSP responded correctly at some point during the MOP sessions and half achieved 100% responding between four and six sessions. Most data in MOP were moderately variable with highly ascending trends and large absolute and relative level changes. The ascending trends are indicative of a facilitative testing effect. This is a significant finding given the widespread use of MOP to assess chained tasks in the literature without little attention to the facilitative testing threats. Although texts and some experimenters discuss such possible concerns, this is the first time that the effects have been experimentally evaluated and repeatedly demonstrated across participants.

The third probe procedure, NOP, resulted in two patterns of responding. Five of the CSP maintained 0% across all six sessions, while the other seven had some responding that ranged between 0-50% with most ranging between 0-17%. The first five participants' data trends are zero-celerating, with no variability or change in level. The participants, who engaged in correct responding, usually allocated correct responses in the first three sessions or last three sessions. Depending on when this occurred, responding affected trend as either decelerating (i.e., Kaylee, Adele, Kassandra, Jonas) or accelerating (i.e., Aricia, Ellie, Elizabeth). For most participants the data were stable, and changes in absolute level were minimal. This more variable responding across participants is interesting, because it demonstrates that the participants had an opportunity to respond but not necessarily learn or maintain responding. Aricia is the exception to this where after completing step four correctly in session four, she maintained it through session six.

Comparatively, SOP resulted in zero-celerating stable data, MOP resulted in accelerating trends with large level changes, and NOP resulted in zero to low levels of responding distributed across different sessions for different participants. Comparing the data for the different probe procedures makes the stated assumptions about possible threats with validity more exaggerated.

For example, the SOP data alone may look as though the CSP were unable to complete the task, but participants responded some in both NOP and MOP when they were given an opportunity. Without comparing MOP to the other procedures, it may be determined that it was the task that lead to the acquisition or some other extraneous variable (e.g., knowing about MOP). By comparing effects, it strongly suggests that MOP have facilitative testing threats because responding was minimal in SOP and NOP and tasks were counterbalanced across participants.

Secondary Student Participants

Dissimilar to the CSP, responding from SSP across all probe procedures was low. One similarity was with SOP, where the SSP data resulted in zero levels of responding for SSP with an overall zero-celerating trend, no variability, or change in level. Like the CSP, it is possible that given the opportunity to complete more than the first step, the participants would have had higher levels of responding. Only four SSP responded correctly at some point during the MOP, but responding never exceeded 17%. This resulted in zero-celerating trends with little to no variability or change in level. Unlike the CSP, there seems to be no indication of a facilitative testing effect. This is noteworthy given the distinct responding patterns with the CSP. Almost all SSP maintained at 0% responding across all sessions in NOP. Two participants responded correctly once, and each of these responses occurred within the first two sessions. Overall the NOP resulted in data that were zero-celerating with almost no variability or change in level. This is somewhat similar to the CSP, but it was less likely for the SSP to correctly respond. When comparing the three probe procedures to each other for SSP, responding was similar. Relatively speaking, most responding occurring with MOP and the least responding with SOP. This rank ordering of level of responding is identical to CSP.

Reasons for the difference in responding on MOP and NOP when comparing the two groups of participants could be attributed to three possibilities: (a) tasks used, (b) learning ability, and (c) differential reinforcement histories. The tasks in this study were nonsense tasks to increase the ability to compare procedures while minimizing threats to internal validity. It is possible that tasks with more functionality would result in similar responding for the SSP. The learning ability of the two sets of participants is important to consider. The CSP group is comprised of students who are either enrolled in graduate school or in their senior year of their undergraduate degree. Comparatively, the SSP group was comprised of individuals with intellectual disability. Given the differential results it is possible that facilitative testing effects with MOP are less likely with individuals with cognitive impairments. Further, there could be an interaction with both learning ability and task type, where more familiar tasks may be more likely to result in a facilitative effect for SSP than nonsense tasks. Lastly, anecdotal information suggests that when compared to SSP, the CSP were more likely to interact with the blocks and attempt to build structures. Therefore, it is possible that the CSP have a greater resistance to extinction resulting in more persistent responding and eventual correct steps. Without attempting the tasks, non-responding for the SSP group resulted in negative reinforcement.

Implications for Research

The results of this study suggest that MOP are likely to lead to facilitative testing threats and SOP have inherent issues with content validity and may lead to inhibitive testing effects. The findings related to MOP are sizable, but only when looking at the data from the CSP. It is recommended that researchers avoid using MOP alone when possible and instead employ other procedures such as combining procedures or exploring new ones. For example a researcher could conduct the first baseline session with a MOP then use NOP for the remainder of sessions.

In this illustration the participant is initially given an opportunity to complete each step to minimize inhibitive testing threats for SOP, while minimizing facilitative testing threats and the reinforcement of odd behaviors from repeated exposures to MOP.

It is possible that facilitative threats are more likely with one population over another or one type of task over another, but those characteristics are yet to be evaluated. If MOP is to be used then consideration should be paid to these variables. One possible way to determine if the task has the possibility of leading to facilitative threats with MOP is to pilot it out with non-participating individuals with similar characteristics. Additionally, if MOP is to be used, more extended baselines (e.g., five versus three) should be used. This recommendation comes from the demonstration of some participants' stable responding in the first three sessions and ascending trends by the fourth session. Figure 10 illustrates this by placing a condition change line between the first and last set of three data points (i.e., between sessions three and four). If the graphs simulated baseline in the first three sessions, it is clear that intervention would have begun with some (i.e., left column in Figure 10) after three sessions. At this point the degree of effectiveness from the intervention alone is unclear. Conversely if the researcher waited five sessions, then the data would have no longer presented stable, possibly leading to other conclusions.

Similarly, SOP should be used with caution and avoided when possible for available alternative options such as combining procedures. For example if a MOP or NOP was first conducted and the participant completed 0% of steps correctly, then assessing in subsequent sessions the researcher could use a SOP. If SOP is used alone, results should be interpreted conservatively as to avoid overly attributing large absolute level changes across conditions to the intervention. Additionally, researchers should identify the use of SOP as a limitation to the content validity of what is being used to sample actual performance in the study.

Given the lack of valid measurement procedures for chained tasks, researchers should continue to evaluate the variables that are more likely to threats with SOP and MOP. The results in this study suggest that threats from using SOP and MOP exist related to measurement, but it is unclear what variables result in the occurrence or nonoccurrence of such effects. Additionally, alternative options for probe procedures such as the NOP or combination of approaches (e.g., one session with MOP, remaining sessions with SOP) should be explored for their validity to assess chained tasks. Although a promising alternative, concerns exist with NOP as well. For example, some CSP in the NOP condition built similar structures from one session to another, suggesting that a chain of errors was reinforced. Other procedural variations may also aide in increasing validity of measurement procedures; such as ensuring the participant understands the task direction or giving more explicit instructions. For example, many of the CSP reported in the social validity questionnaire that they did not know they were supposed to use the grid until later in the sessions. If the task direction was more explicit (e.g., “Create the lifton on the grid” versus “Create the lifton”), more correct responding may have occurred.

The last recommendation is related to the issue with publication bias. Most of the recommendations made for the use of probe procedures come out of suggestions in discussion sections or personal experiences. This is in part from researchers not publishing data, where, for example a participant learned from MOP in baseline and the intervention was unnecessary. The inclusion of such data can only advance the field in further understanding patterns in testing effects from probe procedures.

Limitations

A number of limitations related to this study are important to mention. First, two of the participants were removed from the study after displaying aggressive behaviors. Unfortunately,

these two participants had the same grouping and therefore the absence of their data affected analyzing data in the same task for all probe procedures. It is also possible that if these two participants continued in the study that they would have had some responding, similar to the findings with CSP. Secondly, although careful consideration was paid to creating three equal tasks, there were some patterns in responding by task type. With the CSP and NOP, no responding ever occurred with the ruzzer task and responding occurred most frequently with the lifton task. This was a similar pattern for SSP and NOP where no responding occurred with ruzzer or galtee, but only with the lifton. It is possible that the ruzzer task was more difficult when paired with the NOP while the lifton task was easier. Another pattern also appeared with the SSP and MOP where responding occurred for three out of four participants with the ruzzer, but never for lifton. This pattern latter was not evident with the CSP.

Another limitation comes from the CSP knowing the procedures of the SOP and MOP, therefore affecting their responding. Out of the 12 participants, half answered the question correctly about the two types of probe procedures. Knowledge on the procedures could alert the CSP on how to respond and may result in differentiated responding than those without knowledge. When comparing CSP responding on MOP with knowledge of probe procedures, the only pattern is that the two that did not have an ascending baseline were unable to answer the question. Otherwise, both CSP with and without self-reported SOP and MOP knowledge demonstrated facilitative testing effects. When comparing responders in the NOP procedure to those with and without knowledge of SOP and MOP, there are no discernable differences. This is not surprising given the limited dissemination of the NOP procedure.

Conclusion

This study sought out to answer four research questions related to probe procedures and their possible threats to measurement validity. The study was first conducted with CSP and then replicated with SSP. Findings from the first group of participants suggest that facilitative effects are a real threat with MOP, therefore should be (a) avoided altogether, (b) combined with other procedures, or if to be used (c) include a minimum of five probe sessions prior to intervention. Single opportunity probes have possible inhibitive threats and lack content validity therefore should be (a) avoided if possible, (b) combined with other procedures, and if used (c) include conservative interpretations of results, and (d) acknowledge the use of such procedure as a limitation to content validity. Findings from the second group of participants suggest that although there are threats with the probe procedures, possible variables relating specifically to the participant, learning history, or task may affect the occurrence of such threats with regard to MOP. More research is needed on probe procedures for assessing chained tasks to increase the validity of the measurement procedures being used.

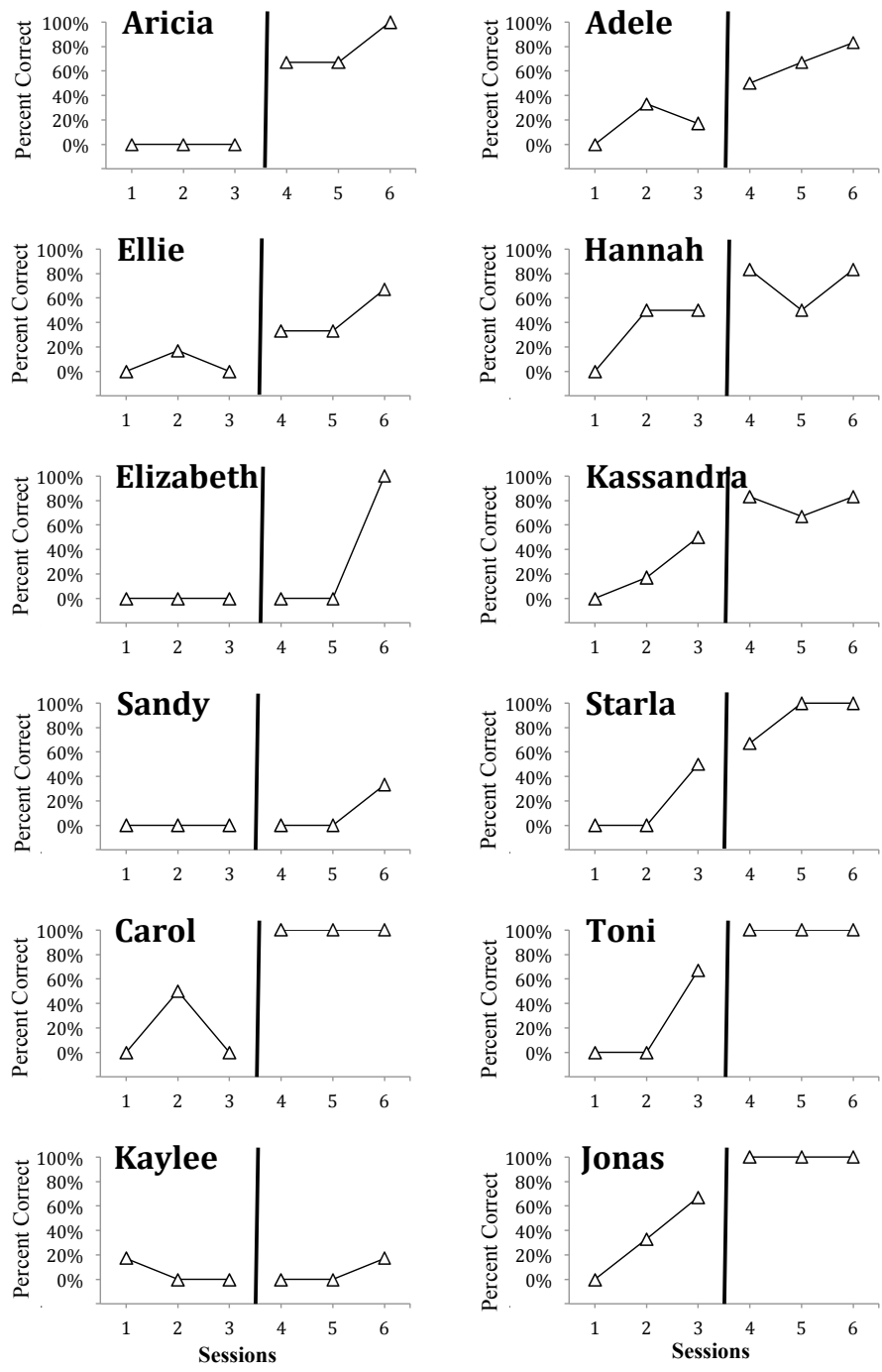


Figure 10. Illustration of facilitative testing threat with CSP inherent with MOP. The left column includes graphs in which the first three data points represent steady state responding. The right column includes graphs in which beginning intervention would be unlikely with unstable baseline data.

REFERENCES

- Alexander, J. L., Ayres, K. M., Smith, K. A., & Shepley, S. B. (in preparation). Evaluating the effects of preference for final products when teaching chained tasks using video modeling.
- Alexander, J. L., Smith, K. A., Mataras, T., Shepley, S. B., & Ayres, K. M. (in press). Potential threats to internal validity in probe procedures for chained tasks: A meta-analysis and systematic review of the literature. *Journal of Special Education*.
- Ayres, K. M., & Ledford, J. R. (2014). Dependent measures and measurement procedures. In D. L. Gast & J. R. Ledford (Eds.) *Single case research methodology: Applications in special education and behavioral sciences* (2nd Ed., pp. 124–153). New York: Routledge Publishers.
- Collins, B. C., Stinson, D. M., & Land, L. A. (1993). A comparison of in vivo and simulation prior to in vivo instruction in teaching generalized safety skills. *Education & Training in Mental Retardation*, 28, 128-142.
- Cook, B. G., & Odom, S. L. (2013). Evidence-based practices and implementation science in special education. *Exceptional Children*, 79, 135-144.
- Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Columbus, OH: Pearson.
- Dunn, L. M., & Dunn, D. M. (2007). *Peabody picture vocabulary test* (4th ed.). Columbus, OH: Pearson.

- Ersoy, G., Tekin-Iftar, E., & Kircaali-Iftar, G. (2009). Effects of antecedent prompt and test procedure on teaching simulated menstrual care skills to females with developmental disabilities. *Education and Training in Developmental Disabilities, 44*, 54-66.
- Farlow, L. J., Loyd, B. H., & Snell, M. E. (1987). Assessing student performance: The effect of procedural contrast between training and probe conditions. Annual Conference of the Association for the Severely Handicapped 14th, Chicago, IL.
- Farlow, L. J., Loyd, B. H., & Snell, M. E. (1988). The implications of the procedural contrast between training and probe conditions on the interpretation of student performance data. Annual conference of the Council for Exceptional Children, Washington, D.C.
- Gast, D. L. (2014). General factors in measurement and evaluation. In D. L. Gast & J. R. Ledford (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (2nd Ed., pp. 85-104). New York: Routledge Publishers.
- Gast, D. L. & Ledford, J. R. (2014). Applied research in education and behavioral sciences. In D. L. Gast & J. R. Ledford (Eds.) *Single case research methodology: Applications in special education and behavioral sciences* (2nd Ed., pp. 1-18). New York: Routledge Publishers.
- Gast, D. L., & Ledford, J. R. (Eds.). (2014). *Single case research: Applications in special education and behavioral sciences*. New York: Routledge Publishers.
- Gast, D. L. & Spriggs, A. D. (2014). Visual analysis of graphic data. In D. L. Gast & J. R. Ledford (Eds.) *Single case research methodology: Applications in special education and behavioral sciences* (2nd Ed., pp. 176-210). New York: Routledge Publishers.
- Godsey, J. R., Schuster, J. W., Lingo, A. S., Collins, B. C., & Kleinert, H. L. (2008). Peer-implemented time delay procedures on the acquisition of chained tasks by students with moderate and severe disabilities. *Education and Training in Developmental Disabilities,*

43, 111-122.

- Hammill, D. D., Pearson, N. A., & Wiederholt, J. L. (2009). *Comprehensive test of nonverbal intelligence* (2nd ed.). Austin, TX: Pro-ed, Inc.
- Hammond, D. L. (2011). *Effectiveness of video modeling delivered via an iPod to teach students with autism to locate library books*. (Unpublished doctoral dissertation). The University of Georgia, Athens, GA.
- Hammond, D. L., Whatley, A. D., Ayres, K. M., & Gast, D. L. (2010). Effectiveness of video modeling to teach iPod use to students with moderate intellectual disabilities. *Education and Training in Developmental Disabilities, 45*, 525-538.
- Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*, 165-179.
- Johnston, J. M., & Pennypacker, H. S. (2008). *Strategies and tactics of behavioral research* (3rd Ed.). New York: Routledge Publishing.
- Kahng, S. Ingvarsson, E. T., Quigg, A. M., Seckinger, K. E., & Teichman, H. M. (2011) Defining and measuring behavior. In Fisher, W. W., Piazza, C. C., & Roane, H. S. (Eds.) *Handbook of applied behavior analysis* (132-150). New York: Guilford Press.
- Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman assessment battery for children* (2nd ed.). Circle Pines, MN: American Guidance System.
- Kazdin, A. E. (2012). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York: Oxford University Press.
- Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston, MA: Allyn & Bacon.

- Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D.M & Shadish, W. R. (2010). Single-case designs technical documentation. Retrieved from: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.
- Moon, M. S., Inge, K. J., Wehman, P., Brooke, V., Barcus, J. M. (1990). *Helping persons with severe mental retardation get and keep employment: Supported employment issues and strategies*. Baltimore: Paul H. Brookes Publishing Co.
- Naglieri, J. A. (2004) *Naglieri nonverbal ability test*. Columbus, OH: Pearson.
- Noell, G. H., Call, N. A., & Ardoin, S. P. (2011). Building complex repertoires from discrete behaviors by establishing stimulus control, behavioral chains, and strategic behavior. In Fisher, W. W., Piazza, C. C., & Roane, H. S. (Eds.) *Handbook of applied behavior analysis*. (250-269). New York: Guilford Press.
- Roid, G. H. (2003). *Stanford-Binet intelligence scales* (5th ed.). Rolling Meadows, IL: Riverside.
- Schuster, J. W., Gast, D. L., Wolery, M. & Gultinan, S. (1988). The effectiveness of a constant time-delay procedure to teach chained responses to adolescents with mental retardation. *Journal of Applied Behavior Analysis*, 21, 169-178.
- Shepley, S. B., Smith, K. A., Ayres, K. M., & Alexander, J. L. (in preparation). The use of video modeling to teach individuals with disabilities to film a video on an iPhone.
- Smith, K. A., Shepley, S. B., Ayres, K. A., & Alexander, J. L. (in preparation). Self-instruction using mobile technology to learn functional skills.
- Smith, K. A., Ayres, K. M., Alexander, J. L., & Mataras, T. K. (2013). The effects of a video prompt embedded in a system of least prompts procedure to teach office skills to individuals with moderate intellectual disability. CEC 2013 Convention and Expo, San Antonio, Texas.

- Smith, K. A., Ayres, K. M., Mechling, L. C., Alexander, J. L., Mataras, T. K., & Shepley, S. B. (2013). The effects of system of least prompts with a video prompt to teach office tasks. *Career Development and Transition for Exceptional Individuals*.
- Snell, M. E., & Brown, F. (2000). *Instruction of students with severe disabilities* (5th ed.). Upper Saddle River, NJ: Pearson Education.
- Tekin-Iftar, E. (2008). Parent delivered community-based instruction with simultaneous prompting for teaching community skills to children with developmental disabilities. *Education and Training in Developmental Disabilities, 43*, 249-265.
- Ventry, I. M. & Schiavetti, N. (1986). Criteria for evaluating research designs. In *Evaluating research in speech pathology and audiology* (72-103) New York: Macmillan Publishing Company.
- Wechsler, D. (1991). *Wechsler intelligence scale for children* (3rd ed.). New York, NY: The Psychological Corporation
- Wechsler, D. (2003). *Wechsler intelligence scale for children* (4th ed.). San Antonio, TX: The Psychological Corporation.
- Wendt, O. & Miller, B. (2012). Quality appraisal of single-subject experimental designs: An overview and comparison of different appraisal tools. *Education and Treatment of Children, 35*, 235-268.
- Williams, G. E. & Cuvo, A. J. (1986). Training apartment upkeep skills to rehabilitation clients: A comparison of task analytic strategies. *Journal of Applied Behavior Analysis, 19*, 39-51.

- Wolery, M. Gast, D. L. & Ledford, J. R. (2014). Comparison designs. In D. L. Gast & J. R. Ledford (Eds.) *Single case research methodology: Applications in special education and behavioral sciences* (2nd Ed., pp. 297-345). New York: Routledge Publishers.
- Wright, T. S., & Wolery, M. (2014). Evaluating the effectiveness of roadside instruction in teaching youth with visual impairments street crossings. *The Journal of Special Education, 48*, 46-58.

APPENDICES

Appendix A

Recruitment Script

I am asking you to take part in a research study. Before you decide to participate in this study, it is important that you understand why the research is being done and what it will involve. This form is designed to give you the information about the study so you can decide whether to be in the study or not. Please take the time to read the following information carefully. Please ask the researcher if there is anything that is not clear or if you need more information. When all your questions have been answered, you can decide if you want to be in the study or not. This process is called “informed consent.” A copy of this form will be given to you.

The purpose of this **research** is to compare different probe procedures commonly used to assess chained tasks. Your participation will involve allowing the researchers to use the information/data collected through your participation to be included in the research. You will be asked to complete three nonsense tasks. You do not have to do anything else. You are being asked to participate because you are an undergraduate or graduate student within the CSSE Department.

If you are interested, you can drop off your signed consent at my office.

Appendix B

Participant Information Sheet

Directions: Please answer all of the questions to the best of your ability.

Name: _____ **Date:** _____

Sex: _____ **Age:** _____

Email Address: _____

Are you currently enrolled as a student at The University of Georgia? _____

Highest level of education (circle one)?

High School Diploma

Master's Degree

Some College

Education Specialist

Associates Degree

Doctorate

Bachelor's Degree

Other (please explain): _____

Degree currently seeking (circle one)?

High School Diploma

Master's Degree

Some College

Education Specialist

Associates Degree

Doctorate

Bachelor's Degree

Other (please explain): _____

What department are you a student of (circle one)?

Special Education

School Psychology

Other (please explain): _____

Are you currently employed? _____

If you are employed, what is your occupation?

Have you ever been employed as a teacher or school psychologist? _____

If you have, please explain.

Do you have a visual impairment? _____

If so, please explain.

What are the two most widely used probe procedures for assessing chained tasks (if unsure, take a guess or answer with “I don’t know”)?

If you are familiar with them, have you used one, both, or neither of the probe procedures in practice or research?

Appendix C

Grid for Probe Sessions

A1	A2	A3	A4
B1	B2	B3	B4
C1	C2	C3	C4
D1	D2	D3	D4





Appendix G

Procedural Fidelity Data Sheet

Participant:								
Date								
Time								
Observer								
Single Opportunity Probe								
Order #								
Set blocks to the left of grid								
Task direction: Create the _____.								
Allow 5s for initiation of steps								
Allow 5s to complete each step								
Provide praise for correct steps								
End session when error occurs by saying, "Stop"								
Provide praise for participating								
Multiple Opportunity Probe								
Order #								
Set blocks to the left of grid								
Task direction: Create the _____.								
Allow 5s for initiation of steps								
Allow 5s to complete each step								
Provide praise for correct steps								
For errors ask participant to close eyes and completes step								
Provide praise for participating								
Natural Opportunity Probe								
Order #								
Set blocks to the left of grid								
Task direction: Create the _____.								
Start timer for 60s								
Provide praise for correct steps								
End session when timer ends by saying, "Stop"								
End session with 30s elapse without correct by saying, "Stop"								
Provide praise for participating								
A. Number of +'s								
B. Number of -'s								
C. Sum A and B								
Percent Correct (A/C*100)								

Appendix H

Screenshots of Social Validity Questionnaire on Google Forms

<p>Probe Study Survey * Required</p> <p>Name * <input type="text"/></p>  <p>Ruzzer Lifton Galtee</p> <p>You completed a total of 6 sessions with all three tasks shown above. Do the best to estimate about how many steps out of 6 you completed correctly and what your typical behavior was. Using the pictures above select the number for questions 1-3.</p>  <p>Ruzzer</p> <p>1. Ruzzer- first session * About how many steps out of 6 did you get correct for Ruzzer on the first session (please select one)?</p> <p>0 1 2 3 4 5 6</p> <p>2. Ruzzer- last session * About how many steps out of 6 did you get correct for Ruzzer on the last session (please select one)?</p> <p>0 1 2 3 4 5 6</p> <p>3. Ruzzer behavior * What did you do for majority of the sessions during Ruzzer?</p> <ul style="list-style-type: none"> <input type="radio"/> Attempted to manipulate blocks <input type="radio"/> Did not attempt to manipulate blocks 	 <p>Lifton</p> <p>4. Lifton- first * About how many steps out of 6 did you get correct for Lifton on the first session (please select one)?</p> <p>0 1 2 3 4 5 6</p> <p>5. Lifton- last * About how many steps out of 6 did you get correct for Lifton on the last session (please select one)?</p> <p>0 1 2 3 4 5 6</p> <p>6. Lifton behavior * What did you do for majority of the sessions during Lifton?</p> <ul style="list-style-type: none"> <input type="radio"/> Attempted to manipulate blocks <input type="radio"/> Did not attempt to manipulate blocks  <p>Galtee</p> <p>7. Galtee- first * About how many steps out of 6 did you get correct for Galtee on the first session (please select one)?</p> <p>0 1 2 3 4 5 6</p> <p>8. Galtee- last * About how many steps out of 6 did you get correct for Galtee on the last session (please select one)?</p> <p>0 1 2 3 4 5 6</p> <p>9. Galtee behavior * What did you do for majority of the sessions during Galtee?</p> <ul style="list-style-type: none"> <input type="radio"/> Attempted to manipulate blocks <input type="radio"/> Did not attempt to manipulate blocks 	<p>10. Out of the three procedures, on which one did you perform the best by the end (i.e., session 6)?</p> <ul style="list-style-type: none"> <input type="radio"/> Ruzzer <input type="radio"/> Lifton <input type="radio"/> Galtee <input type="radio"/> All three about the same <p>11. If you selected one of the three tasks in the question above (i.e., did not select "all three about the same"), why do you think you did the best on the one you selected? *</p> <div style="border: 1px solid black; height: 40px; width: 100%;"></div> <p>12. Did your understanding of what you were supposed to do with the blocks change over time? *</p> <ul style="list-style-type: none"> <input type="radio"/> Yes <input type="radio"/> No <p>12a. If yes, what did you think in the beginning? What did you think in the end?</p> <div style="border: 1px solid black; height: 40px; width: 100%;"></div> <p>12b. If no, what did you think the whole time?</p> <div style="border: 1px solid black; height: 40px; width: 100%;"></div> <p>13. Any other information that you would like to provide, that may be valuable for the purposes of this study?</p> <div style="border: 1px solid black; height: 40px; width: 100%;"></div>
--	---	--