Advances in Neurodevelopmental Disorders (2018) 2:75–85
https://doi.org/10.1007/s41252-017-0049-x

ORIGINAL PAPER

Investigating Generalization Difficulties During Instruction in Language for Learning

Katie Wolfe[1] · Mandy Rispoli[2] · Lacey Taylor[1] · Erik Drasgow[1]

Published online: 16 December 2017

Abstract

Language for Learning is a language curriculum that research supports as effective for teaching language skills to young children with autism spectrum disorder (ASD). However, most of this research has measured direct acquisition and it is unclear to what extent the skills taught in *Language for Learning* generalize beyond the context of the curriculum. Our purpose for this study was to evaluate the effects of *Language for Learning* for producing generalization of labeling skills of two children with ASD when implemented in public school classrooms by the participants' teachers. We used a multiple probe design across language skills to investigate several types of generalization: to untrained visual stimuli, to novel sequences of instructions, and to novel instructors. Results indicate that *Language for Learning* was effective in producing generalization to untrained visual stimuli and to a novel instructor for one skill, but that responding was tightly controlled by the specific sequence of verbal instructions used within the curriculum for other skills. We discuss possible explanations for our findings as well as areas for future research.

Keywords Language · Direct instruction · Language for Learning · Autism · Generalization · Single-case design

Introduction

Individuals with autism spectrum disorder (ASD) often have impairments in receptive and expressive language that can impact their ability to participate fully in homes, schools, and communities (Kjelgaard and Tager-Flusberg 2001; Tager-Flusberg and Joseph 2003). Language deficits can impede the development of social skills and the formation of personal relationships (Garfin and Lord 1986) and can put children at risk for challenging behavior (National Research Council 2001). Research also suggests that receptive and expressive language performance is strongly related to reading comprehension skills (Nation et al. 2006). Nation et al. (2006) found that, among children with ASD with similar levels of decoding skills, oral vocabulary and oral language comprehension were correlated with reading comprehension skills. For example, children with poor reading comprehension skills performed significantly worse on measures of oral vocabulary and oral language comprehension than children with stronger reading comprehension skills. The potential impact of language deficits on other areas of functioning suggests that high-quality language instruction is particularly important for this population (Szatmari et al. 2003).

Language for Learning (LL; Engelmann and Osborn 2008) is a comprehensive language curriculum that has received increasing attention in the research literature because of its potential benefits for children with ASD. *Language for Learning* is a direct instruction (DI) curriculum that focuses on teaching a range of basic and advanced receptive and expressive language skills including labeling objects and actions, using pronouns, labeling prepositions, and using appropriate verb tenses. An overarching goal of the curriculum is to increase the length, complexity, and grammatical accuracy of language. Considering that children with ASD often have deficits in vocabulary, syntax, grammar, and semantics (Condouris et al. 2003; Kjelgaard and Tager-Flusberg 2001) and use fewer words per utterance in natural play contexts compared with typically developing children (Condouris et al. 2003), the skills that are targeted in LL correspond to the aspects of language that may be particularly problematic for this population. Like other DI curricula, LL includes

✉ Katie Wolfe
kmwolfe@mailbox.sc.edu

[1] Department of Educational Studies, University of South Carolina, 820 Main St., Columbia, SC 29208, USA

[2] Department of Educational Studies, Purdue University, 100 N University Street, West Lafayette, IN 47907, USA

effective instructional strategies for students with ASD, such as predictable, fast-paced instruction; many opportunities to respond; and consistent reinforcement and error correction procedures (Watkins et al. 2010).

Several studies that have evaluated the effects of LL on language outcomes for children with ASD have produced promising results (e.g., Flores et al. 2013; Flores and Ganz 2014; Flores et al. 2016; Shillingsburg et al. 2014). For example, Flores and Ganz (2014) compared the effectiveness of LL and discrete trial instruction for teaching language skills to children with ASD and found that participants who received LL performed significantly better on curriculum-based language assessments than those who received discrete trial instruction. However, the research on LL has been limited in terms of its external validity. Specifically, studies conducted to date provide little information about (a) generalization of skills acquired in LL, (b) the effects of LL when implemented by teachers in school-based settings, and (c) the social validity of LL and associated outcomes.

Individuals with ASD often have difficulties with generalization (Schreibman 2005), or the extent to which a skill is performed under conditions that differ from the original teaching conditions (Stokes and Baer 1977). Generalization difficulties may be due in part to the tendency for individuals with ASD to exhibit stimulus overselectivity, which involves responding to a portion of a complex stimulus that is irrelevant to the concept being taught (Ploog 2010). For example, when presented with a picture of a red car and prompted to say, "car," a learner with ASD may respond based on color instead of the entire picture and later may not respond correctly when shown a green car. Stimulus overselectivity makes it particularly important to plan for generalization when designing instruction for students with ASD, and LL contains some features intended to support generalization. Multiple exemplars of visual stimuli are used to decrease the likelihood that students attend to irrelevant features of a stimulus (Stokes and Baer 1977) by varying those stimuli. For example, when targeting the label, "car," LL includes cars of different colors and shapes to teach students that these features are irrelevant to the concept of car and that the relevant features of the concept are those that are common across these exemplars (e.g., four wheels, doors). However, other characteristics of LL may impede generalization. The visual stimuli in LL are restricted to line drawings on a white background, which may limit generalization to other types of visual stimuli (e.g., photographs) because some students with ASD may learn to respond only in the presence of the line drawings. Similarly, the repetitive verbal instructions used in LL may limit generalization to other instructions because some students with ASD may learn to respond only to the specific words or specific sequence of words used in the instructions and not to their overall meaning.

Most studies on LL have relied on the in-program mastery tests as the dependent variable (Wolfe et al. 2017). These tests include the same visual stimuli and sequences of verbal instructions that are used in the curriculum and, as a result, information on generalization of skills acquired in LL is limited. Only one study on LL with children with ASD has systematically evaluated generalization (Wolfe et al. 2017). The authors examined whether three language skills taught using LL generalized to novel visual stimuli (i.e., photographs) and to a novel person and whether the skills were maintained 6–8 weeks following instruction. Both participants accurately performed the targeted expressive language skills, which included labeling objects with a full sentence, answering yes and no, and labeling actions with a full sentence, in the presence of novel visual stimuli and with a novel person. Further, performance of the skills remained above baseline levels 6–8 weeks after the intervention.

Another aspect of generalization that has yet to be examined is the extent to which skills generalize to novel sequences of verbal instructions. *Language for Learning* includes consistent verbal instructions delivered in a specific sequence across lessons to evoke the correct response for a given skill. For example, when asking students to answer yes or no questions about an object, the instructional sequence always starts with labeling the object. For example, while showing a horse, the teacher says, "What is this?" and the students are to respond, "A horse." Then, the teacher asks the series of yes-no questions (e.g., "Is this a boy?", "Is this a cup?", "Is this a horse?"). The first instruction, "What is this?" is not necessary to answer the targeted yes-no questions. However, it is always included in LL exercises that teach this skill, and as a result, some students with ASD may learn to respond accurately to yes-no questions only when they are preceded by an identification question. In natural social situations, students are unlikely to encounter both of these questions—it is more probable that they would be asked one or the other. As a result, the instructions in LL may impair the use of the skill in natural contexts.

Studies of LL conducted with students with ASD to date have also been limited in terms of implementers and settings. According to the U.S. Department of Education (2016), more than half a million students received special education services in 2013–2014 under the category of autism; of these, 90.9% received services within a regular public school (U.S. Department of Education 2015). However, in the studies of LL for students with ASD, the curriculum has been implemented by researchers and research assistants (Flores et al. 2016; Ganz & Flores 2009; Wolfe et al. 2017), student teachers (Flores and Ganz 2014; Flores et al. 2013), or therapists (Shillingsburg et al. 2014), in homes (Wolfe et al. 2017), clinics (Shillingsburg et al. 2014), private schools (Ganz & Flores 2009), and extended school year programs (Flores and Ganz 2014; Flores et al. 2013; Flores et al. 2016). Thus, the studies provide information about the efficacy of the curriculum when researchers are closely involved in its implementation but do not provide information on its effectiveness when used by full-time teachers in public school settings. Effectiveness research, in which a treatment is delivered under real-world conditions by typical implementers, is a critical step in evaluating interventions (Flay 1986; Horner et al. 2005) because variables such as resources, training, treatment acceptability, and supervision may influence outcomes (Detrich et al. 2007).

The purpose of the present study was to evaluate the extent to which language skills acquired by children with ASD in LL generalize and maintain when the curriculum is implemented by full-time teachers in a public school. An additional purpose was to evaluate the teachers' perceptions of the feasibility and effectiveness of LL.

Method

Participants

Two boys with ASD participated in the study. Lyle was 4.5 years old at the beginning of the study and was enrolled in an inclusive preschool class in a public school. On his most recent re-evaluation for special services, he scored 70 on the Preschool Language Scales-5 (age equivalent = 2 years, 10 months; Zimmerman et al. 2011). When prompted, Lyle communicated in single words to request and label items, but rarely used functional vocal communication spontaneously. Lewis was 7 years old at the beginning of the study and was enrolled in a self-contained class for students with moderate to severe disabilities in a public school. Current norm-referenced assessment data were unavailable for Lewis. He requested and labeled items and activities with single words and frequently engaged in echolalia. All diagnoses were conferred by school psychologists and confirmed by the first author.

To be eligible for participation in the study, participants had to have at least one Individualized Education Plan goal pertaining to a language concept addressed in LL, had to place within LL lesson levels based on the placement test, and had to demonstrate the following prerequisite skills: (a) sit and attend for a 5 min, one-on-one instructional session, (b) imitate single words, and (c) imitate basic motor movements. The first author administered the placement test to each potential participant following the protocol in the Teacher Guide. Both participants placed at lesson 1. One potential participant was not eligible because he placed out of the curriculum by scoring 98% on the placement test.

The prerequisite skills were directly assessed by the first author during a 5-min session in which she presented vocal imitation demands (e.g., "say, cat") and motor imitation demands (e.g., "do this" while clapping hands) interspersed with other acquired skills, which were individualized and identified by each participant's teacher. For example, Lewis' acquired skills included labeling numbers and letters presented on flashcards. Participants had to score at least 80% correct on the vocal imitation and motor imitation demands to enroll in the study; both participants met this criterion.

Procedure

Implementers The participants' teachers conducted all remaining study procedures. Lyle's teacher had 3 years of teaching experience at the beginning of the study, was certified in early childhood education, and was pursuing a Masters degree and certification in early childhood special education. Lewis' teacher had 7 years of teaching experience, had a Masters degree in special education, and was certified in severe disabilities.

Neither teacher had experience with DI or LL prior to the study. Both teachers participated in a 3-h training delivered by an instructor who had completed advanced coaching and supervision training from SRA/McGraw-Hill. The training consisted of didactic instruction, practice opportunities, and feedback. Immediately before they began implementing LL, the teachers also participated in two LL fidelity checks in which the teacher delivered LL instruction and the first author and a classroom aide acted as the student. The first author videotaped each fidelity check, recorded data on each teacher behavior specified in procedural fidelity, and provided feedback. The teachers had to score 90% on two consecutive fidelity checks before beginning LL with their students. Lyle's teacher required four fidelity checks to meet this criterion; Lewis' teacher required three.

Setting All sessions took place in public schools in the southeast USA during each participant's scheduled language instruction. Lyle's sessions occurred in a small room (approximately 3 m × 3 m) adjacent to his preschool classroom. The room contained a table, four chairs, and various toys. Lewis' sessions occurred in their teacher's office in their classroom, which contained a desk, a table, and four chairs. Sessions were approximately 20 min in duration and occurred 3 days per week for approximately 7 months.

Materials The teachers used the materials in the LL Teacher Materials Kit, including the Teacher Guide, the Presentation Book, and in-program mastery tests. The teachers also used common three-dimensional objects (e.g., chair, window) in the room as indicated by the curriculum. A token economy system, consisting of a white board with five boxes in which the teacher placed a checkmark, was used during Lewis' LL sessions. Lewis had contacted token economy systems as part of his instruction within his classroom prior to this study. The teacher delivered tokens and allowed Lewis to exchange them for backup reinforcers as described below.

We created an assessment similar to that used in Wolfe et al. (2017) to measure generalization of skills to novel stimuli and to novel sequences of verbal instructions. *Language for Learning* uses colored line drawings on a white background; therefore, we downloaded and printed photographs with a contextualized background as novel visual stimuli and printed the images approximately 8 cm by 8 cm. The LL curriculum uses the same sequence of instructions to evoke the correct response for a given skill, which sometimes includes extraneous questions that are not relevant to the task. For example, when teaching how to label an action with a full sentence, the script always proceeds as follows: "What is this?", "Say the whole thing", "What is this girl doing?", and "Say the whole thing about what this girl is doing." We considered the first two questions in this sequence to be extraneous because they are not directly related to labeling actions with a full sentence. To evaluate generalization to novel sequences of instructions, we omitted the extraneous questions and included only those that were directly relevant to the skill in our assessment. Figure 1 shows sample language skills, target items, visual stimuli, and verbal instructions from the generalization assessment and corresponding items from LL.

Experimental Design We used a multiple probe design (Gast and Ledford 2014) across skill areas and replicated across each participant to evaluate whether LL resulted in generalization to novel visual stimuli and to novel sequences of verbal instructions. Each LL lesson addresses several different skill areas; however, the specific skills were selected in part because they are introduced in a staggered fashion in the curriculum. In other words, one targeted skill was taught while others remained in baseline conditions.

General Procedures We conducted an initial baseline phase, which consisted of probes of each skill conducted by the teacher. We also conducted a probe to measure generalization to a novel person during baseline. When all data were stable, the teacher began implementing LL. These research sessions consisted of a probe followed by instruction in LL. Skills that were not yet taught were in baseline until the lesson introducing them was reached in LL.

Teachers conducted probes using the generalization assessment at the beginning of each session. All skills were included in the probe during the initial baseline sessions and immediately before a new targeted skill was introduced in LL. At other times, the probes included two skills. The skill that was being addressed was always probed, and the others that were either still in baseline or had already been taught were probed on an intermittent basis.

For each participant, the teacher identified 4–5 high-probability instructions that were unrelated to the content taught in LL (e.g., "What color?" or "What number?" while showing a flashcard) and that she delivered before beginning a probe and between skill areas during the probe. She delivered brief verbal praise following correct responses to the high-probability instructions, which were included to enable the participant to contact reinforcement and to maintain responding. No other reinforcement was delivered during the probes of the targeted skill areas. The teacher assessed the skills scheduled for that day by (a) delivering the instruction and holding up the visual stimulus, (b) waiting 5 s for the participant to respond, (c) recording data, and (d) withholding any programmed consequences. The probe procedures were implemented in the same way throughout the duration of the study.

Baseline Prior to LL, the teachers conducted probes of all skill areas for at least five consecutive sessions or until all data paths were stable. During the first three baseline sessions, the teachers also administered one of each of the first three in-program mastery tests to provide a measure of pre-intervention performance. Then the teachers began LL, which

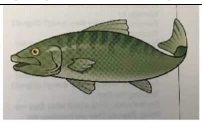
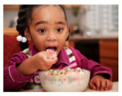
Skill Area		Language for Learning	Generalization Assessment
Yes & No	Visual stimulus		
	Verbal instructions	**1. What is this?** 2. Is this a boy? 3. Is this a fish?	1. Is this a boy? *(No)* 2. Is this a fish? *(Yes)*
Labeling Actions in Pictures	Visual stimulus		
	Verbal instructions	**1. What is this?** **2. Say the whole thing.** 3. What is this girl doing? 4. Say the whole thing about what this girl is doing.	1. What is this girl doing? (*Eating)* 2. Say the whole thing about what this girl is doing. *(This/that/the girl is eating)*

Fig. 1 Sample visual stimuli and verbal instructions from *Language for Learning* (left) and the generalization assessment (right). Instructions in bold were not included in the generalization assessment; responses in italics are sample correct response definitions

targeted the first skill area. The second and third skill areas were in baseline until the participant reached the lesson in LL in which each skill was introduced.

Language for Learning The curriculum was introduced after the initial baseline data were stable across all skill areas. Instruction was delivered in a one-on-one format due to a lack of peers who placed at the same lesson in the curriculum. The teachers implemented the lessons as written in the Presentation Book (Engelmann and Osborn 2008). Each lesson is made up of several exercises, each of which targets a different skill area. For example, one lesson may consist of three exercises, targeting pronouns, personal information, and labeling objects. Each exercise includes a scripted sequence of questions, related visual stimuli, and consistent error correction procedures. The error correction procedure is a model-test-retest format, where the teacher models the correct response, repeats the question, and then intersperses other questions before returning to the missed question to retest. Teachers reinforced correct responding using verbal praise on a variable ratio 2 (VR2) schedule for both participants. The LL curriculum does not direct the teacher to use a specific schedule of reinforcement, so we used a VR2 schedule for praise to maintain a steady rate of responding and to allow for brisk instructional pacing. Lewis' teacher also delivered tokens on a variable ratio 4 (VR4) schedule contingent on correct responding. When Lewis earned five tokens, he exchanged them for a 2-min break that included access to an iPad. Lewis earned between 1 and 3 breaks during each session.

The teachers delivered one lesson per session. Progression through the curriculum was determined by several factors. First, LL specifies that an exercise should be repeated until students can respond accurately and independently to all instructions within that exercise (Engelmann and Osborn 2008). Sometimes, a participant required several days of repeating an exercise to meet this criterion. Second, teachers conducted the in-program mastery tests included in the curriculum every 10 lessons. If the participant's accuracy is below 90% on this test, LL directs the teacher to repeat specific exercises and then repeat the in-program mastery test. Third, when a participant was scheduled to begin the first lesson targeting the next skill area, we examined the probe data for the current skill to evaluate whether it evidenced a visually apparent change from baseline. If it did, the teacher began the LL lesson targeting the next skill area. If the data for the current skill had not changed, we made modifications to LL lessons targeting that skill and continued probing prior to moving onto the next lesson (see following description of procedures). Total session duration, including probes and LL lessons, was on average approximately 20 min. The mean duration of LL lessons, minus breaks, was 9.2 min.

Curricular Modifications For both participants, at least one skill area did not generalize after at least 4–6 sessions targeting those skills, despite accurate performance in the lessons and on the in-program mastery tests. Therefore, we modified the sequence of instructions in the LL exercises targeting the skill to reflect the sequence of instructions on our generalization assessment; that is, we eliminated the extraneous questions that preceded the questions that were directly relevant to the skill area. In the curricular modification phase, we repeated the most recently completed LL lesson in its entirety (i.e., all exercises) until the participant's probe data demonstrated a change before progressing to the next LL lesson.

Generalization to a Novel Person To assess generalization to an unfamiliar person, a trained research assistant assessed all skill areas once during the initial baseline and once on the final day of instruction using the generalization assessment.

Maintenance Three weeks after the final lesson was delivered, teachers evaluated maintenance of all skill areas by administering the generalization assessment using the same procedures used throughout the study.

Measures

We measured the percent of correct responses on probes of targeted language skills using the generalization assessment as the primary dependent variable. Lyle's target skills were labeling objects (skill 1; lesson 1), labeling actions (skill 2; lesson 17) and labeling prepositions (skill 3; lesson 28). Lewis' target skills were labeling objects (skill 1; lesson 1), answering yes and no (skill 2; lesson 15), and labeling actions (skill 3; lesson 17). For each skill, we selected two to four targets taught in the first three LL lessons addressing that skill. For example, the targets for labeling objects with a full sentence were fish, banana, car, and chair.

All skills contained six to eight trials, and each question was asked once during a probe. Across all questions and probe types, a correct response was counted if the participant emitted a correct response (see Fig. 1) within 5 s of the instruction. Sometimes, our defined correct response included more grammatically correct variations (e.g., a sentence can begin with "this," "that," or "the") than accepted by the curriculum or in-program mastery tests (e.g., sentence must begin with "this") to improve the validity of the assessment (complete assessment and response definitions for all participants are available from the first author). Incorrect responses included any vocalization other than a defined correct response, self-corrects, or no response within 5 s. To obtain the percent correct for each skill, the number of correct responses was divided by the total number of trials for that skill and multiplied by 100.

We also measured three secondary dependent variables. First, using the generalization assessment, we evaluated generalization to a novel person before and after LL. Second, also using the generalization assessment, we measured maintenance 3 weeks after the last instructional session. Third, to measure direct acquisition, the teachers also conducted the in-program mastery tests, which occur every 10 lessons, contain the same visual and verbal stimuli as the curriculum, and measure all language skills taught up to the mastery test. The teachers conducted each mastery test at least twice: once during the initial baseline period and again when it was reached in the curriculum. If the participant did not meet the 90% criterion on the mastery test, the teacher repeated specified exercises and then repeated the mastery test, per LL. The curriculum identifies the specific correct response for each of 25–30 trials (e.g., "Touching my nose"), and when scoring the in-program mastery tests, we evaluated participants' responses against this standard. A global percent correct for each mastery test was calculated by dividing the number of correct responses by the total number of trials and multiplying by 100.

Interobserver Agreement To measure the reliability of the dependent variables, a trained research assistant collected data from audio recordings on 33% of probes across all phases of the study. The teachers audio-recorded all probes, and the first author randomly selected 33% of them for reliability assessment. The first author collected data in vivo on an additional 28% of probes across all phases of the study and on all in-program mastery tests. In both cases, the secondary data collector (i.e., the first author or research assistant) marked each participant response as correct or incorrect, and the teacher's data was compared to the secondary data collector's data on a point-by-point basis (Kazdin 2011). Mean agreement for Lyle's probes for skills in baseline was 94% (range = 88–100%) and in intervention was 93.6% (range = 87.6–100%). Mean agreement for probes with Lewis for skills in baseline was 98% (range = 92–100%) and in intervention was 95% (range = 89.4–100%).

Procedural Fidelity A trained research assistant recorded procedural fidelity data on 33% of probes across all phases and 37% of LL lessons using audio recordings of both types of sessions. The teachers audio-recorded all probe and LL lessons, and the first author randomly selected 33 and 37% of them for procedural fidelity checks. For probes, procedural fidelity was measured by recording whether the teacher correctly performed each of the following on each trial: (a) delivers instruction as indicated, (b) waits 5 s, and (c) does not deliver any consequence. We also recorded whether she delivered high-probability instructions. The number of correct behaviors was divided by the total number of possible behaviors and multiplied by 100. Across all participants and both teachers, the mean probe fidelity was 97% (range = 92–100%).

Fidelity of LL was evaluated for each instructional trial within lessons by recording whether the teacher followed the script as written and used the correct error correction procedure following an error. The error correction procedure consisted of the following: (a) modeling the correct answer, (b) repeating the missed question, and (c) returning to the beginning of the exercise to repeat preceding questions and retest the missed question. All three steps had to be completed correctly for the error correction procedure to be scored as correct. We divided the number of correct behaviors by the total number of opportunities and multiplied by 100. Across both teachers and all participants, mean accuracy for following the script was 99% (range = 89–100%) and for following the error correction procedure was 88% (range = 84–100%). Omitting the retest was the most common error in implementing the procedure. (Participant- and teacher-specific fidelity data is available from the first author).

Social Validity The teachers completed a short, online social validity survey approximately 1 month following the end of the study. Using a 5-point Likert-type scale, ranging from strongly disagree to strongly agree, they rated whether (a) LL was worth the time and effort, (b) LL was easy to implement, (c) they would use LL with other, similar students, (d) they would modify LL if needed, (e) they would recommend LL to other teachers, and (f) LL produced meaningful changes in their students' language. They also had the opportunity to provide open-ended responses about the curriculum and the study.

Data Analyses

We primarily used visual analysis to analyze the data according to recommendations by Kratochwill et al. (2013). Specifically, we evaluated the level, trend, and variability within each phase and compared these same features between adjacent phases to determine whether a change in the dependent variable was present (i.e., a basic effect). To determine whether a functional relation was present, we examined whether there were three basic effects within one or both participants' multiple probe designs (Kratochwill et al. 2013). We also conducted a vertical analysis at each phase change to evaluate the continued baseline stability of skills not yet taught in LL. To supplement visual analysis, we calculated means for each skill in each phase.

Results

Lyle Figure 2 contains the results of the probes on the generalization assessment for Lyle, who placed at lesson 1 and for whom we selected labeling objects, labeling actions, and labeling prepositions as targeted skills 1, 2, and 3, respectively.

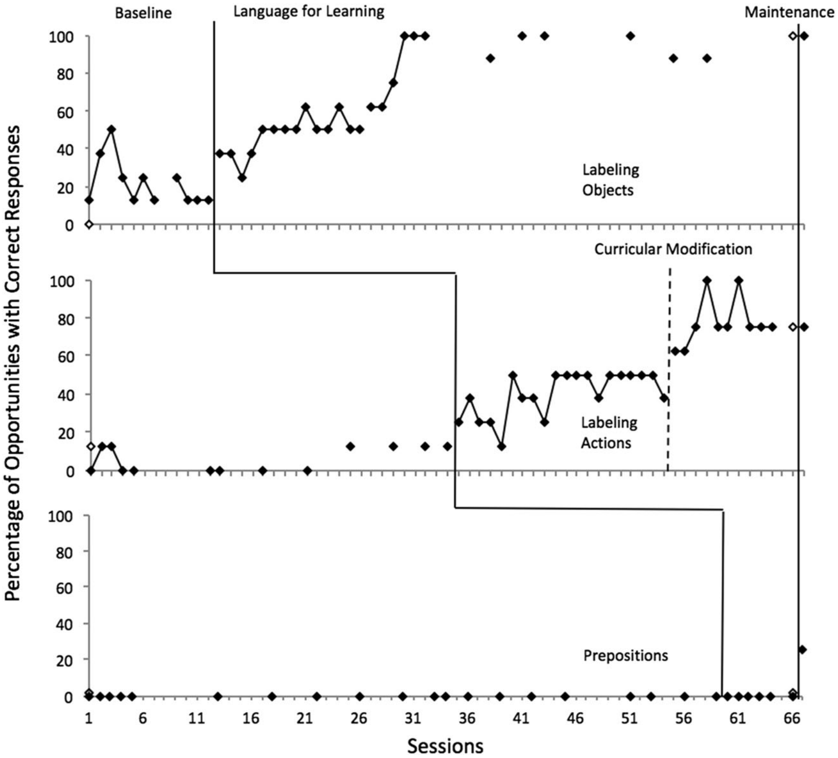


Fig. 2 Lyle's performance on generalization probes. Open diamonds represent generalization to a novel person

In baseline, Lyle's performance of skill 1 was variable before stabilizing between 12.5 and 25% correct ($M = 22\%$), performance of skill 2 varied between 0 and 12.5% correct ($M = 6\%$), and performance of skill 3 was stable at 0% correct ($M = 0\%$). When LL was introduced to target skill 1, there was an immediate increase in the level and trend of Lyle's performance on skill 1 ($M = 63\%$) but probes of skills 2 and 3, which remained in baseline, indicate that those skills remained stable. Lyle reached lesson 17 in *Language for Learning*, where skill 2 is introduced, after session 34. Like skill 1, there was an immediate change in the level and trend of Lyle's performance on the generalization assessment for skill 2 ($M = 40\%$); however, his performance stabilized at 50% correct on skill 2 after 11 lessons addressing this skill. After session 54, Lyle was performing the skill independently and with 100% accuracy during LL lessons and on the in-program mastery tests. Therefore, we modified the LL lessons as previously described to eliminate the extraneous verbal instructions included in the sequence teaching the skill. Immediately following this modification, Lyle's performance on probes of skill 2 increased to between 75 and 100% ($M = 78\%$). The teacher then progressed to lesson 28, where skill 3 is introduced. Lyle's performance of skill 3 remained at 0% for several sessions, and the school year ended before we could introduce modifications.

Lyle's teacher conducted the first three in-program mastery tests once during baseline and again when they were reached in the curriculum following lesson 10 (session 21), lesson 20 (session 42), and lesson 30 (session 64). Lyle's scores on those tests in baseline, prior to LL instruction, were 33, 24, and 16%, respectively, and during instruction were 96, 92, and 93%, respectively. We also measured generalization to a novel instructor and maintenance after LL was terminated, both using the generalization assessment. Lyle's results indicate that skills 1 and 2 generalized to a novel instructor and maintained 2 weeks following the end of LL.

Lewis Figure 3 depicts the results of the probes on the generalization assessment for Lewis, who also placed at lesson 1. We selected labeling objects, answering yes and no, and labeling actions as skills 1, 2, and 3, respectively, for Lewis. In baseline, Lewis' performance of all skills was generally stable around 0% correct ($M = 1$, 0, and 3.5%, respectively). Several lessons targeting skill 1 were conducted before the skill began to generalize, and a steep upward trend is apparent following session 17 ($M = 70\%$). Skills 2 and 3, which remained in baseline, remained stable at 0% correct on intermittent probes. After session 27, Lewis reached lesson 15, where skill 2 is introduced. There was no change in Lewis' performance on the generalization assessment of skill 2 after several sessions of LL lessons targeting this skill ($M = 4\%$); however, after six sessions, we noted that he was performing the skill independently and with 100% accuracy during LL lessons. As with Lyle, we modified the LL lessons to eliminate the extraneous verbal instructions in the sequence teaching the skill. Following just two sessions with this modification, Lewis' performance on probes of skill 2 immediately increased 100% ($M = 80\%$). The pattern for skill 3 is similar; after

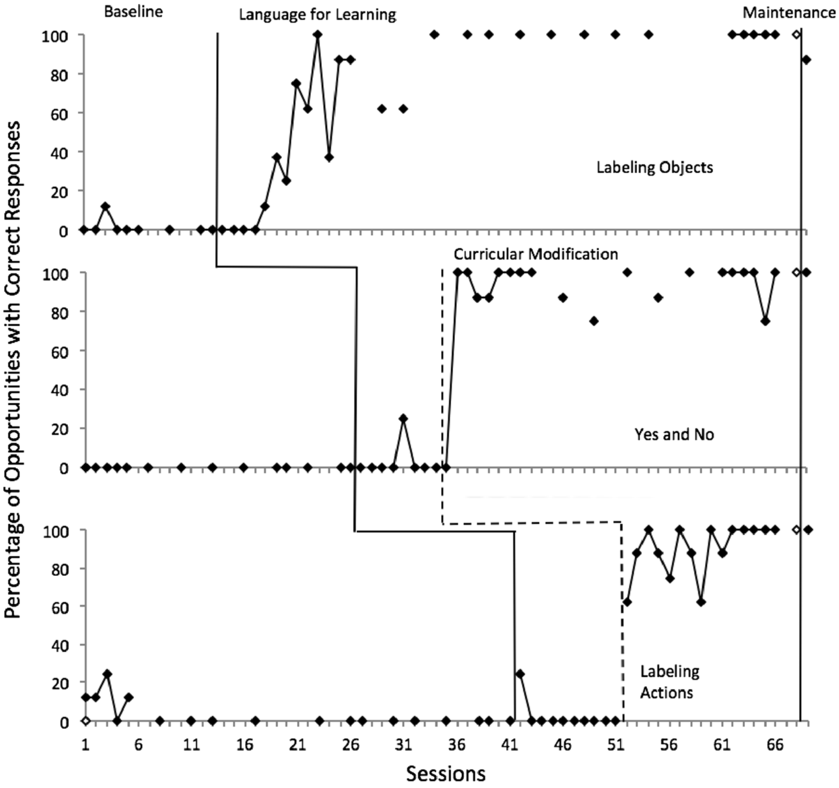


Fig. 3 Lewis' performance on generalization probes. Open diamonds represent generalization to a novel person

several sessions of LL without progress on the generalization assessment ($M = 4\%$), we observed that Lewis was performing skill 3 accurately and independently during LL lessons. We modified the sequence of instructions in LL and noted an immediate increase to 60% accuracy after just one lesson. His performance of this skill was variable before stabilizing at 100% for several sessions ($M = 90\%$).

Lewis' baseline performance on the first three in-program mastery tests was 24, 10, and 8% correct. When his teacher conducted them again following lesson 10 (session 26), lesson 20 (session 44), and lesson 30 (session 66), his percent of correct responses were 80, 81, and 83%, respectively. Following each mastery test, his teacher repeated the indicated exercises, and on the following day, repeated the mastery test. Lewis' scores on the second attempt at each mastery test were 90, 92, and 91%, respectively. All three skills generalized to a novel instructor and maintained 3 weeks following the end of instruction in LL.

Social Validity The teachers rated their agreement with the six statements about LL on a 5-point Likert-type scale. Both teachers agreed or strongly agreed that LL was worth the time and was easy to implement, and that they were likely to use it in the future and to recommend that other teachers use it. However, they were both neutral about whether it produced meaningful changes in their students' language skills. On open-ended questions, both teachers reported that they liked the easy-to-follow instructions in the curriculum, but one noted concerns about the functional use of the skills, and the other indicated that the error correction procedure was time-consuming and that she would likely modify it if she were to use LL again.

Discussion

Our primary purpose in conducting this study was to examine the effectiveness of LL for producing several types of generalization when implemented by teachers in public school classrooms with children with ASD. Our findings indicate that, although both participants acquired the language skills as taught in LL per the mastery test results, only one skill generalized without curricular modifications for each participant. Thus, we were unable to demonstrate a functional relation between LL and generalization for either participant. However, we implemented curricular modifications when skills did not generalize, and in all cases, accurate performance on probes rapidly increased. The effects of the modification were demonstrated with one skill with Lyle and across two skills with Lewis. All skills that were performed with accuracy also generalized to a novel instructor and maintained 3 weeks following instruction.

In a previous study evaluating generalization and LL, the targeted language skills of two participants with ASD generalized to novel visual stimuli and a novel person (Wolfe et al. 2017). In the current study, only one of three skills generalized for each participant, despite accurate performance on that skill during LL lessons and on the in-program mastery tests. Interestingly, both participants generalized the same skill, labeling objects, following LL instruction. One potential explanation for the generalization of this skill, and not others, may be the participants' learning histories with respect to each skill. It is likely that the participants had contacted language instruction focused on labeling nouns, as this is a common skill addressed in early intervention programs. However, they may not have experienced instruction addressing the other skills included in the study (i.e., labeling yes/no, labeling actions, labeling prepositions).

The generalization assessment in this study included two types of generalization: novel stimuli and novel sequences of verbal instructions. The curricular modifications consisted of manipulating the sequence of instructions while holding the visual stimuli constant, and therefore enabled us to identify which type of generalization was impacting performance on the probes. If the participant accurately performed the skill following the modification, it would suggest that the sequence of instructions in the curriculum was exerting strong control over responding and hindering generalization to the probes. On the other hand, if the participant was still unable to perform accurately on the probes with the modified instructions in LL, it might suggest that the visual stimuli were inhibiting generalization because the participant was directly taught to respond to the novel sequence of instructions through the modifications.

In all three cases in which the curricular modification was implemented (i.e., skill 2 for Lyle and skills 2 and 3 for Lewis), probe performance improved immediately and dramatically within 1–2 sessions. As a result, our data suggest that skills failed to generalize because of the specific sequence of verbal instructions in LL. A potential explanation for this finding relates to stimulus overselectivity (Ploog 2010). It is possible that participants were attending to an irrelevant aspect of the stimulus (i.e., the extraneous questions at the beginning of the instructional sequence) and the curricular modifications produced attending to the relevant aspects of the stimulus (i.e., the questions pertaining directly to the targeted skill). Alternatively, it is possible that the sequence of verbal instructions formed a chain (Cooper et al. 2007), wherein the participants could not perform the terminal responses (i.e., the generalization assessment) in the absence of the initial "links" (i.e., the extraneous questions and associated responses in LL). The curricular modification, which resulted in immediate improvement on the probes, may have broken this chain by removing the irrelevant verbal instructions from LL.

The research design and the curricular modifications that enabled us to isolate the variable impeding generalization could serve as a model for future research on generalization. Repeated measurement of generalization as a primary dependent variable allows the researcher to systematically manipulate components of the instruction when a skill fails to generalize. Such manipulations can identify the specific feature(s) of instruction that are contributing to restricted stimulus control and may result in modifications that improve generalization outcomes.

Although conclusions are tentative until replicated by other researchers and with other participants, our results suggest that some students with ASD may not be able to use the skills that they learn in LL unless the exact sequence of verbal instructions used in the curriculum is always used to evoke the response. As the sequence of instructions is unlikely to be used outside the context of LL, this finding has significant clinical implications. When implementing LL with students with ASD, researchers and practitioners should specifically monitor generalization while measuring acquisition using the in-program mastery tests. If skills are not generalizing, LL instruction may be adjusted to omit irrelevant questions or vary the sequence of instructions (i.e., train loosely, Stokes and Baer 1977).

In contrast to previous studies of LL for this population, in the present study, the curriculum was implemented by full-time teachers in a public school classroom over the duration of a school year. Procedural fidelity data indicate that the teachers implemented the curriculum with high integrity, and the participants' direct acquisition data from mastery tests replicate the results of other studies in which the curriculum was implemented by researchers or therapists in home or clinic settings (e.g., Shillingsburg et al. 2014; Flores et al. 2016).

The social validity data also represents a novel contribution of this study. Though only two teachers participated, the results of the social validity questionnaire indicate that they believed the curriculum to be easy to implement and worth the time, and that they would consider using LL for students with similar needs in the future. However, both teachers were neutral about the social significance of the changes in receptive and expressive language that they observed and noted that if they used the curriculum again, they would likely modify it to meet their students' needs. They also both expressed concerns about generalization of skills given the repetitive nature of the curriculum. Although the teachers' perception of LL was generally positive, it is difficult to draw strong conclusions about the social validity of the curriculum based on such a limited sample. Researchers in applied behavior analysis and special education have highlighted the importance of considering the acceptability of goals, interventions, and outcomes (Horner et al. 2005; Wolf 1978), and future research on LL should incorporate measures of social validity to inform the evaluation of the curriculum.

Limitations and Future Research Directions

The results of this study should be considered within the context of several limitations. First, the curriculum was delivered in a one-to-one format because of a lack of peers in each classroom placing at the same lesson as the participants. Much previous research on LL has also been conducted in a one-to-one instructional arrangement, but it is designed to be delivered in a group format and should be evaluated as such. Future research could examine whether generalization of skills acquired in LL is improved when participants contact the instruction in a small group compared to a one-to-one format. Additionally, receiving highly structured instruction in a group format may result in improvements in group instruction skills in general that may benefit children with ASD. For example, children with ASD may acquire ancillary skills including hand raising, turn taking, and choral responding through group-delivered LL instruction that may generalize to other group instructional contexts.

Second, the generalization assessment was a relatively proximal measure considering that it included (a) a small number of target items for each skill area, (b) only one visual stimulus per target item, and (c) some of the same verbal instructions that are included in the curriculum. Although the generalization assessment was somewhat restricted, it is important to note that both participants had difficulty generalizing at least one skill even on a proximal measure. Future research on LL should include more distal measures of generalization. For example, researchers could examine generalization to novel instructions (e.g., “Use a full sentence” instead of “Say the whole thing”), and related curricular modifications to produce looser stimulus control (i.e., train loosely, Stokes and Baer 1977). Natural language samples could provide more information about the functional use of skills acquired in LL. Standardized measures of receptive and expressive language could provide information about specific gains in various aspects of language (e.g., vocabulary, syntax, and grammar).

Third, the researcher conducted the placement test with both participants at the beginning of the study. This may limit external validity because the teachers did not implement all procedures involved in the use of the LL curriculum. Further, it is possible that the participants would have performed differently on the placement test had it been administered by a familiar person. In future studies that evaluate the effects of LL when delivered by teachers or other school-based staff, it will be important for those individuals to be responsible for all procedures related to the implementation of the curriculum.

In conclusion, researchers and practitioners should continue to examine the practical significance of outcomes associated with LL. Although our assessment evaluated several types of stimulus generalization and represents an extension of previous research, it was conducted under contrived testing conditions. The ultimate determinant of the utility of LL for children with ASD should be the functional and appropriate use of the skills in natural situations (e.g., play, social, academic).

Author Contributions KW: conceptualized the study, conducted the study, and wrote the manuscript. MR: collaborated in the design, writing, and editing of the manuscript. LT: collected interobserver agreement and procedural fidelity data. ED: collaborated in writing and editing of the final manuscript.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethics Statement This study was reviewed and approved by the University of South Carolina Institutional Review Board.

References

Condouris, K., Meyer, E., & Tager-Flusberg, H. (2003). The relationship between standardized measures of language and measures of spontaneous speech in children with autism. *American Journal of Speech-Language Pathology, 12*(3), 349–358.

Cooper, J. O., Heron, T. E., & Heward, W. E. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River: Pearson Merrill Prentice Hall.

Detrich, R., Keyworth, R., & States, J. (2007). A roadmap to evidence-based education: building an evidence-based culture. *Journal of Evidence-Based Practices for Schools, 8*(1), 26.

Engelmann, S., & Osborn, J. (2008). *Language for learning*. Columbus, OH: Science Research Associates.

Flay, B. R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive Medicine, 15*(5), 451–474.

Flores, M. M., & Ganz, J. B. (2014). Comparison of direct instruction and discrete trial teaching on the curriculum-based assessment of language performance of students with autism. *Exceptionality, 22*(4), 191–204.

Flores, M. M., Schweck, K. B., & Hinton, V. (2016). Teaching language skills to preschool students with developmental delays and autism spectrum disorder using Language for Learning. *Rural Special Education Quarterly, 35*(1), 3.

Flores, M. M., Nelson, C., Hinton, V., Franklin, T. M., Strozier, S. D., Terry, L., & Franklin, S. (2013). Teaching reading comprehension and language skills to students with autism spectrum disorders and developmental disabilities using direct instruction. *Education and Training in Autism and Developmental Disabilities, 48*(1), 41–48.

Ganz, J. B., & Flores, M. M. (2009). The effectiveness of direct instruction for teaching language to children with autism spectrum disorders: Identifying materials. *Journal of Autism and Developmental Disorders, 39*(1), 75–83.

Garfin, D. G., & Lord, C. (1986). Communication as a social problem in autism. In E. Schopler & G. B. Mesibov (Eds.), *Social behavior in autism* (pp. 133–152). New York: Plenum Press.

Gast, D. L., & Ledford, J. R. (Eds.). (2014). *Single case research methodology: applications in special education and behavioral sciences*. New York, NY: Routledge.

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*(2), 165–179.

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings*. Oxford: Oxford University Press.

Kjelgaard, M. M., & Tager-Flusberg, H. (2001). An investigation of language impairment in autism: implications for genetic subgroups. *Language and Cognitive Processes, 16*(2–3), 287–308.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34*(1), 26–38.

Nation, K., Clarke, P., Wright, B., & Williams, C. (2006). Patterns of reading ability in children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 36*(7), 911–919.

National Research Council. (2001). *Educating children with autism*. Washington, DC: National Academies Press.

Ploog, B. O. (2010). Stimulus overselectivity four decades later: a review of the literature and its implications for current research in autism spectrum disorder. *Journal of Autism and Developmental Disorders, 40*(11), 1332–1349.

Schreibman, L. (2005). *The science and fiction of autism*. Cambridge, MA: Harvard University Press.

Shillingsburg, M. A., Bowen, C. N., Peterman, R. K., & Gayman, M. D. (2014). Effectiveness of the direct instruction language for learning curriculum among children diagnosed with autism spectrum disorder. *Focus on Autism and Other Developmental Disabilities, 30*(1): 44–56. https://doi.org/10.1177/1088357614532498.

Stokes, T. F., & Baer, D. M. (1977). An implicit technology of generalization. *Journal of Applied Behavior Analysis, 10*(2), 349–367.

Szatmari, P., Bryson, S. E., Boyle, M. H., Streiner, D. L., & Duku, E. (2003). Predictors of outcome among high functioning children with autism and Asperger syndrome. *Journal of Child Psychology and Psychiatry, 44*, 520–528.

Tager-Flusberg, H., & Joseph, R. M. (2003). Identifying neurocognitive phenotypes in autism. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 358*(1430), 303–314.

U.S. Department of Education, National Center for Education Statistics. (2016). Digest of Education Statistics, 2015 (NCES 2016–014). Retrieved from https://nces.ed.gov/programs/digest/d15/tables/dt15_204.30.asp?referrer=report

U.S. Department of Education, Office of Special Education Programs (2015). Individuals with Disabilities Education Act (IDEA) database. Retrieved from http://www2.ed.gov/programs/osepidea/618-data/state-level-data-files/index.html#bcc.

Watkins, C. L., Slocum, T. A., & Spencer, T. D. (2010). Direct instruction: relevance and applications to behavioral autism treatment. In E. A. Mayville & J. A. Mulick (Eds.), *Behavioral foundations of effective autism treatment* (pp. 297–319). Cornwall-on-Hudson, NY: Sloan Publishing.

Wolf, M. M. (1978). Social validity: the case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis, 11*(2), 203–214.

Wolfe, K., Blankenship, A., & Rispoli, M. (2017). Generalization of skills acquired in language for learning by young children with autism spectrum disorder. *Journal of Developmental and Physical Disabilities*, 1–16.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2011). *Preschool language scales* (5th ed.). San Antonio, TX: Psychological Corporation.