Teaching Science Skills and Knowledge to Students with Developmental Disabilities:

A Systematic Review

Abstract

A comprehensive review of the literature was conducted to identify current practice on teaching science to students with Intellectual Disability (ID) and/or Autism Spectrum Disorder (ASD) in relation to two review questions – students' science outcomes and students' and teachers' experiences of the interventions. Six databases related to education, psychology and science were systematically searched. A detailed protocol can be viewed on PROSPERO (registration number – 42017057323). Thirty studies were identified that reported on science interventions and 20 on student/teacher experiences of the interventions. The majority of the studies targeted science vocabulary and concepts. Other targets included inquiry skills and comprehension skills. The majority of the interventions used components of systematic instruction (n=23). Five studies focused on self-directed learning and two on comprehension based instruction. Students and teachers reported positive experiences of the interventions. The findings suggest that components of systematic instruction in particular might be effective in teaching science content to students with ID and/or ASD. Further research is needed to explore the effectiveness of identified interventions on teaching more complex science skills and with students with severe disabilities. Some limitations related to the search strategy are highlighted.

Keywords*:* science education, science curriculum, developmental disabilities, intellectual disability, autism spectrum disorder, special educational needs

Learning science is a core entitlement for students during their compulsory years of schooling in many parts of the world. It not only enhances learners' curiosity and understanding of the world around them (Browder & Spooner, 2011), but it also provides students with an important set of inquiry skills to help them evaluate evidence and ideas The functional application of science content can also provide the basis of employment for some students with disabilities (Collins, Terrell, & Test, 2017; Rizzo, & Taylor, 2016), as well as deepening their understanding of their own bodies, weather changes, and the natural world. Furthermore, skills acquired during science lessons can help students with disabilities access instruction alongside their peers in general education classrooms and learn essential life skills (Spooner, Knight, Browder, Jimenez, & DiBiase, 2011). It is, therefore, essential from a rights and a functional perspective that we provide meaningful access to the science curriculum to all students, including students with disabilities.

Like other contemporary education practice, science education has moved away from a standard model of schooling focused on learning facts to a pedagogy that aims to promote a deeper understanding of key science concepts (Sawyer, 2008). An example of this move is embodied within the *Next Generation Science Standards (NGSS)* a set of core values adopted by states to improve the provision for teaching scientific content and skills through more practical scientific experiences (National Research Council, 2013). More recent thinking in science education has also moved towards the development of a balanced science curriculum based on 'big ideas' in science, aimed at promoting science as an interesting and relevant subject that is central in the creation of ethically aware and critically informed young people (Harlen, 2015). It is within this context that science remains a important subject to be understood by all students regardless of gender, culture, ethnicity or disability.

How students learn science

McGinnis and Kahn (2014) report four main perspectives on learning that have shaped current thinking on teaching science to students with special needs (including intellectual disability and autism spectrum disorder): developmental (the thinking of children and adults is different and it changes throughout the life), behavioral (learning is the result of connection between stimuli and behavior and it continues until prescribed mastery criteria are reached), sociocultural (individuals' development is a results of interactions between multiple factors such as culture, environment, etc.), and cognitive (focus is placed on mental processes, such as memory, perception, attention and metacognition).

According to McGinnis and Kahn (2014), many practitioners favoring a cognitive (constructivist) perspective employ teaching approaches that enable students to build their understanding of scientific ideas by undertaking practical scientific inquiry tasks (often called inquiry-based learning), whereas those preferring the behavioral model place a greater emphasis on teaching more knowledge-based learning programmes aimed at attaining mastery of predetermined learning objectives. Inquiry-based learning based on the principles of cognitive science is commonly referreed to as constructivism where learners contruct their own understanding of conepts and ideas from minimal information (Steffe & Gale, 1995; Kirschner et al., 2006), It is important to note, however, that the term constructivism in science education refers to a theory of learning rather than a clearly defined theory of teaching. In practice, the division between the behavioral and constructivist approaches is often more nuanced than the binary division commonly outlined in the literature.

Of more practical significance than the discussion on *how* students learn science is the distinction between how students learn science (i.e. inquiry-based learning) and their ability to *work* scientifically (i.e. undertaking the process of science inquiry where learners apply their science knowledge and skills to answer questions). The ultimate goal of teaching science is to equip students with the knowledge and skills to enable them to carry out the process of science inquiry to answer testable questions and/or gather information in a systematic manner. Whether learners have acquired the necessary science skills through inquiry-based or direct teaching approaches is perhaps of secondary importance to the main goal of ensuring students are able to apply these skills to enable them work scientifically.

More recent research has tried to draw together findings from cognitive and developmental psychology to describe a set of core skills that underpin children's early learning in science (Tolmie, 2016). The proposed core components of initial science learning for young children are: (i) accurate observation, (ii) the ability to extract and reason explicitly about causal connections, and (iii) knowledge of mechanisms that explain these comnnections. This work details the important part language acquisition and group work play in supporting children's emergent scientific ideas, especially for the skills of predicting and reasoning associated with casual observations.

Science and the rights of students with disabilities

In the USA, 13% of all school age children have disabilities (Snyder, de Brey, & Dillow, 2018). In England, 14.4% of all students are characterized as having special educational needs (Department for Education, 2017). Despite students with disabilities being a significant minority in the school age population, they are still under-represented in research studies in the field of education, especially students with more severe disabilities (Spooner & Browder, 2015). McGinnis and Kahn (2014) report that there is also an over-representation of students from 'racial' and ethnic minorities among students with disabilities or special educational needs which might be related to poverty, students' academic achievement being devalued, and language use (e.g., with multiple languages being spoken at home).

Internationally, policy and guidance is clear about the inclusion of students with disabilities in science education. The No Child Left Behind Act (2002) emphasized schools' obligation to provide high quality education to all students and required schools in the USA to assess all students' progress in reading, mathematics and science. The Every Student Succeeds Act (2015) shifted accountability to individual States and left much more flexibility to how students' knowledge is being assessed while continuing to emphasize the use of evidence-based practice in teaching students with disabilities. UNESCO's Education 2030 agenda envisions inclusion of all historically excluded pupils, including those with disabilities, by 2030 together with the creation of more safe and accessible educational establishments (United Nations Educational, Scientific and Cultural Organization, 2017). The central point of the international agenda is the right of every learner to equal access to education. In the UK, under the Equality Act 2010, schools have an obligation to provide access to education to all students and make reasonable adjustments for students with disabilities (Department for Education, 2014a). Teaching should be personalized to ensure meaningful access to the curriculum for all students. Teachers are encouraged to frequently assess students' progress and set goals that are achievable yet ambitious (Department for Education, 2015). Moreover, The Special Educational Needs and Disability code of practice (Department for Education, 2015) recommends the choice of teaching approaches based on available evidence.

Teaching Science

In the late 1990s, the United States National Science Education Standards (NSES) shifted attention to the use of inquiry-based instruction (learning focused on students posing questions, exploring and testing ideas to enable them to construct their own understanding) and emphasized that "learning science is an active process" (National Research Coucil, 1996, p. 20). The NSES requires science education to cover eight standards including: science concepts; science inquiry; physical, life and earth and space science; science technology, history of science and social and personal perspectives on science (National Research Council, 1996). The Next Generation Science Standards (NGSS) identify core standards within all grades on three dimensions: core ideas, practice and crosscutting concepts (National Research Council, 2013). The standards focus on the development of students' comprehension of key science concepts and processes as well as their ability to develop and test hypotheses and evaluate evidence.

In England, science education standards are organized based on age-related key stages focused on three basic aims: the development of science knowledge and concepts, and scientific inquiry skills ('working scientifically') (Department for Education, 2014c). Schools are required to teach students science across all ages. However, mainstream content can often be inaccessible for students with developmental disabilities (Spooner, McKissick, Knight, & Walker, 2014), where the teaching paradigm is often focused on inquiry or discovery-based learning. These strategies are often successful with typically developing learners in mainstream settings but can be less effective for less able students and students with disabilities (Rizzo, & Taylor, 2016).

Previous research on science education and students with developmental disabilities

In the present review, we focused on science education for students with intellectual disability (ID) and/or autism spectrum disorder (ASD) – describing these groups of children with the general term developmental disability (DD). "Intellectual Disability (intellectual developmental disorder) is a disorder with onset during the developmental period that includes both intellectual and adaptive functioning deficits in conceptual, social, and practical domains." (American Psychiatric Association, 2013). Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by deficits in social communication, social interaction and repetitive behaviors (American Psychiatric Association, 2013). Those two populations were chosen due to similarity of challenges that the learners face and the relatively limited existing research literature and guidelines for professionals. Findings for both “diagnoses” are clearly differentiated in the current review to help practitioners find relevant information in relation to their population of interest.

Students with disabilities or special educational needs (SEN) have poor attainment in science. For example, in England only 24% of students aged 4-7 with SEN achieved the expected attainment level in science (Department for Education, 2014b). According to educational progress data published in the USA by National Assessment of Educational Progress (NAEP), in 2015 students with disabilities in grade 4 (9-10 years old), grade 8 (13-14 years old) and grade 12 (17-18 years old) achieved scores between 124 and 131 (out of 300) in science in comparison to scores between 153 and 158 for students with no disabilities (The Nation’s Report Card, n.d.). Given the cognitive difficulties associated with ID in particular, the science attainment gap is likely to be much larger for children with DD, although specific data on these disability groups are not available at national levels.

Three previous systematic reviews have been published on science education for students with various DDs. Courtade, Spooner and Browder (2007) focused on research published between 1985 and 2005 on teaching science concepts to students with significant cognitive disabilities. The search strategy was based on seven science standards from NSES and included a systematic literature search of two databases. Eleven studies, all using single-case experimental designs, were identified. The most recent included study was published in 2003. The total of students in all included studies was 58. All interventions used components of systematic instruction – an approach focused on teaching observable and measurable behaviors and promoting generalization (Browder & Spooner, 2011) (see later for definition). Courtade et al. (2007) concluded that students with significant cognitive disabilities can benefit from teaching strategies like time delay, modelling, and errorless learning to acquire science skills and that a strong emphasis should be put on generalization of learning.

Spooner et al. (2011) focused on research literature published between 1985 and 2009 on science education for students with severe DDs. The conceptual framework of science education used in the review was developed after consultations with experts in the field of science education and severe disabilities and the search strategy was based on eight science standards from NSES. Five databases were searched. Seventeen studies were included in the review, of which 14 were rated as being of adequate or high quality. Spooner et al. (2011) concluded that systematic instruction is an evidence-based practice for teaching science to students with DDs. Spooner et al. (2011) also emphasized that most recent research suggests that students with severe disabilities can successfully learn science skills based on the general curriculum.

Rizzo and Taylor (2016) analyzed literature on inquiry-based instruction for students with various disabilities. Three databases were searched. Twelve studies published between 1992 and 2013 were included, and the authors concluded that students' science achievement improved when inquiry teaching techniques were used, but that it is not an effective teaching strategy on its own. Rizzo and Taylor (2016) also concluded that students with disabilities require support to access inquiry-based instruction and that their science gains increase when components of explicit instruction are used.

The most recent systematic review on all components of science education (Spooner et al., 2011) included articles published prior to 2009. Since then, new articles have been published on teaching science to students with various DDs, thus an updated review is warranted. Moreover, none of the previous reviews focused on the entire population of student with ID and/or ASD. Courtade et al. (2007) and Spooner et al. (2011) focused on students with severe ID only (IQ below 55) and Rizzo and Taylor (2016) focused on all students with disabilities. Spooner, McKissick and Knight (2017) in their summative paper on evidence-based practices for students with severe disabilities reported that at the time of the last comprehensive systematic review on teaching science to students with DD (Spooner at al., 2011) published studies were mainly focused on traditional functional curriculum domains (e.g. safety skills). Since then, more research targeting skills that are part of National Curriculum in the UK or National Science Education Standards and The Next Generation Science Standards in the US have been published. Additionally, in the last two decades a shift in science education has taken place from a more knowledge-based curricula to more creative methods of teaching that encourage deeper understanding (Sawyer, 2008 National Research Council, 2013). This is reflected in the number of studies published in recent years on science education for mainstream populations, as well as for students with disabilities. Due to those dynamic changes in the field and the shift in the understanding of science education, a new systematic review is warranted. An additional aim of the present systematic review was to extend the findings of Spooner et al. (2011) by including students' and teachers' experiences of the interventions. These data are crucial to fully understand effectiveness and feasibility of different interventions.

The current review focused on the following questions: What interventions have been developed to teach science skills and knowledge to children with developmental disabilities (DD)?, and What are the views and experiences of students with DDs and their teachers on interventions used to teach science?

Method

The protocol for this systematic review was registered on PROSPERO (https://www.crd.york.ac.uk/prospero/ - also available from the corresponding author on request) before any searches started (registration number – 42017057323) to enhance transparency and rigor. PRIMSA guidelines (Moher, Liberati, Tetzlaff, & Altman, 2009) for reporting systematic reviews were used in the current paper.

Review focus and inclusion criteria

This review focused on research evaluating educational interventions for teaching science to students with DD. The population of interest included children and young adults up to 25 years old with an ID and/or ASD. Participants had to have one or both diagnoses to meet the inclusion criteria. Science education was defined in line with UK standards as "scientific knowledge and conceptual understanding through the specific disciplines of biology, chemistry and physics" (Department for Education, 2014c, p. 168) and understood as "the pursuit and application of knowledge and understanding of the natural and social world following a systematic methodology based on evidence" (Science Council, n.d., para. 1). The contexts of interest were individual or group settings in schools or further education colleges (including international equivalents). Studies describing interventions delivered in different settings were excluded. Included studies reported outcomes from interventions compared against teaching as usual (typical lessons as per students' timetables) or other interventions. Studies with no comparison but reporting change from baseline measures were also included. Included studies had to report either students' change in science skills and knowledge (review question 1) or students' and teachers' opinions and experiences of the science intervention effectiveness, usefulness, or ease of use (review question 2). For review question 1, any quantitative research with a comparison design was included (e.g., controlled trials, single group pre-test post-test designs, and single-case experimental designs). For review question 2, any quantitative or qualitative studies reporting data on students' and/or teachers' opinions or experiences of the science intervention were included. Studies could be included with mixed samples of students with different disability diagnoses or no disability as long as the data on students with DD were reported separately.

Search strategy

Six databases were searched in March 2017. Databases were chosen based on their area of focus related to education, psychology and science. In August 2017, forward and backward reference searches of all included studies (and the Spooner et al., 2011 review) were conducted. Following that, five active researchers in the field of science education for students with DDs whose studies had been identified were contacted to enquire about any relevant unpublished research. Forward and backward searches were completed for any newly identified studies until no new studies were identified.

The search strategy was developed based on the terms related to science education, ID and ASD with a help of a University based librarian and applied in the following databases: ERIC, Education Research Complete, PsycINFO, Social Science Citation Index, British Education Index, and ASSIA. Search terms were organized into two lists – one containing terms related to ID and ASD and the second terms related to science education (see Table S1). Due to the nature of science education, the search strategy was deliberately designed to be wide to minimize the chance of potentially relevant studies being missed. Search terms within each list were separated with “OR” and Lists 1 and 2 were combined with “AND”. All terms were searched in titles, keywords and abstracts.

The review focused only on research papers published in English and Polish – the languages in which the research team were competent. No restrictions regarding publication date were applied. Additionally, database searches were limited to peer review journal articles only.

Study selection

After the relevant articles were identified in the databases, all results were exported to an electronic data program and scanned for internal and external duplicates. Following that, the first author scanned the titles, abstracts and keywords of all the results against the inclusion/exclusion criteria. At this stage, articles were excluded only if they clearly did not meet the review criteria. To examine reliability of this selection, the fifth author independently scanned 20% of randomly selected results against the inclusion/exclusion criteria. Agreement was calculated by dividing the number of agreements by the number of agreements and disagreements and multiplying by 100%. Reliability for initial study selection was 99.85% (kappa = .93). Full text versions of all studies identified at initial screening were obtained, and a checklist of all inclusion/exclusion criteria was used to establish whether to include papers in the review. Agreement for this full selection stage was 96.62% (Kappa = .88). Inclusion disagreements were discussed with a third research team member for resolution.

Quality appraisal and data extraction

After all the articles were screened, quality appraisal tools were applied to the included articles by the first author. Appropriate tools were chosen depending on each study's design. The Critical Appraisal Skills Programme (CASP) checklist for randomized controlled trials (CASP, 2017) was used for studies incorporating randomized controlled trial (RCT) designs. This checklist consists of 11 questions and is divided into three sections in relation to results – their validity, their value, and if they can be helpful in practice. The same checklist, excluding the randomization question, was used for the non-randomized controlled studies. For parts A and C of the checklist, each question is assigned either yes or no answer based on the information provided in the article. For part B, appropriate information from the results section of the paper is provided. For articles using single-case experimental designs the Quality Indicators tool developed by Horner et al. (2005) was used. This tool consists of 21 indicators within seven main sections: participants and setting, dependent and independent variables, baseline, internal and external validity, and social validity. Each indicator is assigned either yes or no answer based on information provided in the article, and a quality appraisal score is derived from the total number of quality indicators present.

Data extraction used a piloted bespoke tool for this review that included the following information: author, year, origin, population characteristics, setting characteristics, study characteristics, intervention characteristics, intervention delivery characteristics, quantitative outcomes, together with data on participants' and teachers' experiences of the intervention. The first author completed the data extraction for all included articles while the fifth author independently completed extraction for 20% of randomly selected articles. Studies included in the systematic review were summarized using narrative synthesis.

Results

Study selection

Figure 1 summarizes the study selection process. 27,205 records were identified through initial database searches and 28 through reference searches. No additional studies were identified through contact with active researchers in the field. After removal of 7,233 internal and external duplicates, the initial screening of titles, abstract and keywords led to the exclusion of 19,817 records. Subsequently, full texts of 183 studies were assessed for eligibility. From these, 151 records were excluded with the main reasons recorded, and full text copies of two articles could not be obtained. Quality appraisal and data extraction was completed for the remaining 30 articles.

Study characteristics

The included studies were published between 2003 and 2017 with the majority of the studies published in or after 2010 (n=22). Of the 30 included studies, 29 were from the USA and one from the UK. All 30 studies included data on students' science related learning (research question 1). Twelve studies included multiple educational outcomes but the current review reports only on students with science related targets. Twenty studies reported students' and teachers' experience and opinions on science interventions (research question 2). Tables 1, 2 and 3 present a summary of 30 studies included in the systematic review.

Participants. The mean number of participants with science targets reported across all included studies was 3.9 (range 1-21), with most of the studies reporting outcomes for three students (n=14). In total, 118 students were involved in the included studies.

Facilitators. Seventeen studies included interventions delivered by school staff – either general or special education teachers or paraprofessionals (e.g., Karl, Collins, Hager, & Jones Ault, 2013; Knight, Creech-Galloway, Karl, & Collins, 2017; Riesen, McDonnell, Johnson, Polychronis, & Jameson, 2003). Seven interventions were implemented by researchers (e.g., McMahon, Cihak, Wright, & Bell, 2016), three by peer tutors (e.g., Hudson, Browder, & Jimenez, 2014) and one by researchers and school staff (Roberts & Joiner, 2007). Two articles did not contain clear descriptions about intervention facilitators (e.g., Miller, Doughty, & Krockover, 2015).

Setting. All 30 studies were conducted in school or college settings. Fifteen interventions were delivered in students' typical classrooms (special education classroom, resource rooms or self-contained classrooms) (e.g., Riggs, Collins, Kleinert, & Knight, 2013; Miller et al., 2015; Smith, Spooner, Jimenez, & Browder, 2013b). Ten studies included interventions delivered in general education classrooms (e.g., McDonnell et al., 2006). Two interventions were delivered in both special and general education classrooms (e.g., Collins, Evans, Creech-Galloway, Karl, & Miller, 2007) and another two in different settings – one in a kitchenette (Miller & Taber-Doughty, 2014) and one in a greenhouse (Collins et al., 2017).

One study did not provide a detailed description of the setting (Carnahan & Williamson, 2013).

Design. Twenty-eight studies incorporated single-case experimental designs (e.g., Jimenez, Lo, & Saunders, 2014; Karl et al., 2013**;** Riggs et al., 2013; Smith et al., 2013b) and two used group designs (Browder et al., 2010; Roberts & Joiner, 2007).

Science targets. The majority of the studies targeted science vocabulary and concepts (n=18) (e.g., Collins et al., 2007; Knight, Smith, Spooner, & Browder, 2012). Two studies focused on science inquiry skills (e.g., Miller & Taber-Doughty, 2014) and six studies included targets related to both, science inquiry and vocabulary (e.g., Jimenez, Browder, & Courtade, 2009). Two studies focused on textbook comprehension (e.g., Carnahan & Williamson, 2013), while the remaining two focused on listening comprehension of science content (Hudson et al., 2014) and chemical and physical properties (Collins, Hager, & Creech-Galloway, 2011).

Interventions. The majority of interventions used components of systematic instruction (see later for definition) (n=23) (e.g., Browder et al., 2010; Collins et al., 2007; Knight, Spooner, Browder, Smith, & Wood, 2013; McDonnell et al., 2006). Five studies used self-directed learning (see later for definition) (e.g., Roberts & Joiner, 2007) and two studies focused on comprehension based instruction (see later for definition) (e.g., Carnahan and Williamson, 2013). The seven studies where the main intervention components were based on systematic instruction also contained elements of different teaching approaches – peer tutoring (n=3), technology based instruction (n=3) and self-directed learning (n=1). Three studies that used self-directed learning also incorporated different approaches – task analysis (n=2) and technology based instruction (n=2).

Generalization and maintenance. Fifteen studies assessed generalization of targeted skills beyond the teaching context (e.g., Riggs et al., 2013; Heinrich, Collins, Knight, & Spriggs, 2016) and 15 did not (e.g. Johnson, McDonnell, Holzwarth, & Hunter, 2004; Knight, Wood, Spooner, Browder, & O'Brien, 2014). Twenty articles included data on maintenance of skills over time (e.g., Riggs et al., 2013; Smith et al., 2013b) and 10 did not (e.g., McMahon et al., 2016; Miller & Taber-Doughty, 2014).

Perceptions and experiences of the interventions - participants. The majority of studies that reported data on participants' opinions and experiences of the intervention focused on both students and teachers (n=5) (e.g., Carnahan, Williamson, Birri, Swoboda, & Snyder, 2016; Jimenez et al., 2009) or students only (n=5) (e.g., McMahon et al., 2016). Five studies reported only perceptions of teachers (e.g., Carnahan & Williamson, 2013) and two of students, peer tutors and teachers (e.g., Jimenez, Browder, Spooner, & DiBiase, 2012). The remaining three studies reported experiences of peer tutors and teachers (n=2) (Hudson et al., 2014) and parents and teachers (n=1) (Courtade, Browder, Spooner, & DiBiase, 2010).

Perceptions and experiences of the interventions – tools. Fifteen studies incorporated a single tool to gather data on experiences and perceptions of the intervention (e.g., Johnson et al., 2004) and five used multiple tools (e.g., Jimenez et al., 2012). Ten studies used questions with rating scales (e.g., Miller & Taber-Doughty, 2014). Seven studies incorporated tools with a mixture of open- and close-ended questions (e.g., McMahon et al., 2016) and six used surveys with closed-ended questions (e.g., Smith, Spooner, & Wood, 2013a). The remaining four studies used open-ended questions (n=2) (Agran, Cavin, Wehmeyer, & Palmer, 2006), focus groups (n=1) (Jimenez et al., 2012) and incidental observations reported by school staff (n=1) (Agran et al., 2006).

Synthesis

Systematic instruction. Systematic instruction is "teaching focused on specific, measurable responses that may either be discrete (singular) or a response chain (e.g., task analysis), and that are established through the use of defined methods of prompting and feedback based on the principles and research of applied behavior analysis" (ABA) (Browder, 2001, p.95). It focuses on five components: socially important skills, operationally defined targets, data collection to monitor progress, stimulus control transfer methods and generalization (Browder & Spooner, 2011). Spooner and Browder (2015) described systematic instruction as one of three most significant advances for students with severe disabilities. Systematic instruction has been used to teach a range of skills from functional living skills like cooking (Mechling, Gast, & Fields, 2008) to navigating around the community (e.g. Taber, Alberto, Hughes, & Seltzer, 2002), and teaching academics (e.g. Knight et al., 2013). While a range of different systematic instruction teaching methods can be used to teach different skills, educators generally apply four steps to implement the instruction. These steps start with (1) defining target skills, then move to (2) planning and defining instructional methods, next they (3) implement the intervention, and later (4) assess students' progress and modify the methods if needed (Browder & Spooner, 2011).

Twenty-three studies that used systematic instruction to teach students science content were included (see Table 1 and S4). Twenty-two of them used single-case experimental designs and one used a group design. The interventions used procedures such as: task analysis (breaking down a complex task into smaller steps); embedded instruction (providing instruction for target skills during on-going activities); constant time delay (procedure involving delivery of the prompt after a specific amount of time after the instruction, usually starting at zero seconds and systematically increasing the interval); simultaneous prompting (the prompt is delivered straight after the instruction and then gradually faded out; controlling probes are conducted before the training to determine if the skills has been acquired); system of least to most prompts (hierarchy of prompts used to help the students, starting from the least intrusive); scripted lessons (an instructional strategy that provides teachers with scripts with exact information on how to teach each target and deliver the instruction) and explicit instruction (an active teaching method involving modelling). Simultaneous prompting procedures and embedded instruction were the two most frequently used teaching approaches. Fourteen interventions were delivered by school staff (either a teacher or paraprofessional), six by researchers, and three by peer tutors. Three studies also used computer-assisted instruction (CAI) – a teaching approach involving the use of different means of technology to deliver the instruction.

Additionally, two studies evaluated effectiveness of a science curriculum for students with DD. Jimenez et al. (2014) taught three students with moderate to severe ID and ASD science vocabulary and concepts using scripted lessons with and without guided notes. Two students made good progress after the intervention was implemented and one student made little progress. Smith et al. (2013b) taught three students with severe disabilities science vocabulary and concepts during inquiry-based lessons using systematic instruction. All the students made good progress when the intervention was implemented.

Students in all studies showed increases in dependent variables as a result of intervention implementation. However, some students did not reach mastery criterion. For example, Collins et al. (2017) used a simultaneous prompting procedure to teach science content related to photosynthesis embedded in a practical skill (plant care) to four students with ID. The rate of correct responses for all students improved at post-test compared to pre-test. However, none of the students met the mastery criterion before receiving three additional training sessions.

Three students in three of the studies showed no or very little increase in target skills. For example, Fetko, Collins, Hager and Spriggs (2013) used a simultaneous prompting procedure to teach three students with ID and ASD science vocabulary embedded in a leisure activity training (UNO game). The rate of correct responses increased from 0% at pre-test to 100% at post-test for two students but the third student did not show any progress. There were no studies where none of the students showed an increase in the dependent variable when the intervention was implemented, perhaps due to publication bias.

Fourteen studies reported students' and/or teachers' experiences and opinions on the systematic instruction intervention used, four studies also included peer tutors' views, and one study included parents' views. Overall, reported experiences of the interventions were positive with students reporting that the intervention was enjoyable, and they would like to try it again in the future. Teachers reported that the intervention targeted skills important for their students and was effective in improving their science outcomes and feasible to implement. Attitude surveys conducted with peer tutors showed increases in their positive attitudes towards students with disabilities. Parents indicated that they thought that it was important that their children could access science lessons. They also reported increased interest in science skills of their children.

Out of 23 studies using systematic instruction methodology, only ten reported students' ethnic/'racial' background and only two reported their primary language (see Table 1). Available data suggested a lack of diversity. The majority of the students were African American (n=27), Caucasian/White (n=13) or Hispanic (n=5). Two studies reported students' primary language as English. The remaining 13 studies did not provide any information about ethnic/'racial' background of the participants.

Table 1 summarizes quality appraisal results for the systematic instruction studies (see Table S2 for more details). Ten studies met all 21 indicators and were categorized as high quality; nine met 20 indicators, and the remaining three studies met 19 indicators. The main area of weakness for the single-case experimental design studies was the description of participants (n= 6): although these articles provided a general description of participants, they failed to include detailed information about participant's primary diagnosis. A further area of weakness was the lack of an operational description of the dependent variable (n=5). Three studies did not meet the magnitude of change criteria as some participants in those studies made no or very little progress after the intervention was implemented. Since 11 studies targeted multiple skills, including other areas of education apart from science, a second quality assessment was conducted using the same tool (Horner et al., 2005) with the focus on science targets only. Seven articles received the same quality score during the revised quality appraisal when only science related intervention was evaluated. In contrast, four articles received a lower score. Most of those discrepancies were due to design limitations. Overall, the quality appraisal results were relatively unaffected by this sensitivity analysis adjusted to focus on science aspects only.

One RCT study (Browder et al., 2010) was high quality except for whether participants and staff were blinded to the intervention, although this would not be feasible to achieve in the school context (see Tables 1 and S3).

Self-directed learning. Self-directed "strategies allow students to manage, direct, and regulate their own learning and permit students to plan, execute, and evaluate actions based on problem solving and self-directed decision making" (Agran et al., 2006; p.231). This type of instruction allows students to take control over their learning (Browder & Spooner, 2011).

Five studies used self-directed learning to teach science to students with DD (see Table 2 and S5). Four studies used single-case experimental design and one used a group design. Some of the interventions included: a self-determined learning model of instruction (instructional model that teaches students to set goals, implement curriculum augmentation strategies and self-monitor progress; Agran et al., 2006), augmented reality application (digital tool that blends the physical environment with digital content; McMahon et al., 2015), a self-monitoring checklist, and concept mapping (method of constructing visual maps to help establish connections between different concepts; Roberts & Joiner, 2007). One intervention was delivered by school staff, one by the researcher, and one by both a researcher and teacher. Two articles did not provide detailed descriptions of intervention facilitators. Two interventions used CAI and three incorporated systematic instruction components: task analyses, and exemplar and non-exemplar training.

Students in all studies showed increases in the level of the dependent variable as a result of the intervention implementation. For example, Miller and Taber-Doughty (2014) used a self-monitoring checklist and science notebooks to teach inquiry skills to three students with ID. All students showed a large increase in the rate of correct responses after the intervention was implemented compared to baseline. Moreover, their rate of responding remained high during generalization probes.

Four studies also reported students' experiences and opinions of the interventions. Overall, students expressed positive experiences indicating that they enjoyed the interventions and they helped them learn science. None of the studies reported teachers' experiences.

Of five studies using self-directed learning, only two reported students' ethnic/'racial' background and none reported their primary language (see Table 2). Available data suggest that samples were not diverse. The majority of the students were Caucasian (n=4) and two participants were labelled as Latino. The remaining three studies did not provide any information about ethnic/‘racial’ background of the participants.

Table 2 summarizes the quality appraisal results for self-directed learning studies (see Table S2 for more details). Five articles used single-case experimental designs. Three studies met all 21 Horner et al. (2005) indicators and are categorized as high quality (e.g., Miller & Taber-Doughty, 2014). One article met 20 indicators (Agran et al., 2006) since no information about procedural fidelity was included. Since one study (Agran et al., 2006) targeted multiple skills, including other areas of education apart from science, a second quality assessment was conducted with the focus on science targets only. The article received the same quality score again suggesting that the quality of the study was not affected by including multiple targets.

One study (Roberts & Joiner, 2007) used a within-participant crossover experimental design and the quality was assessed using the CASP form for RCTs (Critical Appraisal Skills Programme, 2017) without the randomization question. The results are presented in Table 2 (see Table S3 for more details). The study was high quality.

Comprehension based instruction. The “goal for comprehension instruction is for students to learn to transfer skills acquired in reading narrative texts to comprehending the elements in expository texts” (Browder & Spooner, 2011; p.143). The narrative texts include novels, shorts stories and similar, whereas expository texts include, for example, textbooks.

Two studies used comprehension based intervention to teach science to students with DD (see Table 3 and S6). Both used single-case experimental design and the intervention was delivered by the school staff. The interventions included a compare-contrast strategy package (intervention including contrasting and comparing signal words and summarizing information; Carnahan & Williamson, 2013) and multicomponent text structure intervention (intervention pack involving instruction in different types of text patterns; Carnahan et al., 2016). Students in both studies showed increases in the level of the dependent variable as a result of the intervention implementation. For example, Carnahan and Williamson (2013) used a compare-contrast strategy package to teach science textbook comprehension to three students with ASD. The rate of responding of all students was already quite high at baseline, but students made progress when the intervention was implemented and maintained their responding over time.

Both studies also reported students' and teachers' experience and views on the interventions. Students reported that the intervention helped them learn textbook comprehension and they would like to continue using it. The teachers indicated that the interventions were feasible, targeted skills important for their students and were effective in teaching them new skills. Neither study reported students' ethnic/'racial' background.

Table 3 summarizes quality appraisal results (see Table S2 for more details). Both studies met all 21 Horner et al. (2005) indicators and were categorized as high quality.

Discussion

The main aims of the review were to identify what methods had been reported in the education literature to teach aspects of science to students with developmental disabilities, and, for the first time, to report on students' and other stakeholders perceptions and experiences of these interventions. We begin our discussion by briefly summarizing the main findings from our review. Finally, we discuss our findings within the conceptual framework of the main theories of learning in science education, and describe how more systematic approaches to teaching can be used to help teach students with DD to work scientifically.

Spooner et al. (2011) concluded that systematic instruction was an evidence-based practice for teaching science to students with moderate to severe disabilities. Although the the current review also identified systematic instruction research, we identified additional teaching approaches (self-directed instruction and comprehension based instruction) that might also be effective in teaching specific science content to students with DD.

Three main teaching approaches were identified in this systematic review. The majority of the studies (n=23) used systematic instruction. Of 90 participants, only three students did not make progress in their target skills as a result of the intervention. Although this may represent a reporting bias, these data on progress in outcomes are consistent with Spooner et al.'s (2011) conclusion that systematic instruction is an effective teaching technology for science education for students with DD. In addition, teaching strategies and targeted outcomes were very diverse in the current review such that quantitative synthesis of the studies was not possible. Thus, any conclusion about the effectiveness of systematic instruction should be made with caution. For the first time, we also reported data on stakeholders' experiences and the 10 studies reported participants' perceptions indicated that systematic instruction interventions were valued and feasible to implement.

Multiple teaching methods were often combined in one intervention. However, simultaneous prompting procedures and embedded instruction were the two most frequently used teaching approaches. The majority of the interventions were implemented by school staff in the general education classrooms or students' typical classrooms. The students' experiences were positive, and teachers commented that the targeted skills were socially important. Additionally, quality appraisal results indicate that the majority of the studies using systematic intervention (n=19) were of high or acceptable quality with only four studies obtaining a lower rating. One study using a group design was also of adequate quality. Overall, systematic instruction seems to be a promising approach to teach science to students with DD. However, more high-quality research is needed, especially using randomized controlled trial designs (RCT), to establish its effectiveness for students with severe DD. More high quality single case experimental design research is also warranted, especially studies sharing procedures and outcomes that can later be synthesized quantitatively

The second teaching approach identified was self-directed instruction. Five studies that used this method reported positive outcomes for students with DD. All 22 participants made progress in their target skills and students in four of the studies reported positive experiences of the intervention (no data reported on experiences of students in one study). Teachers' opinions and perceptions were not reported. Quality appraisal results indicate that four studies were of high or adequate quality. One study using a group design was also of acceptable quality. Self-directed instruction seems to be a promising approach to teach science, especially inquiry skills, to students with DD. More high-quality research is needed to establish its effectiveness across a variety of outcome measures. Again, the variability in teaching approaches and outcomes precluded a quantitative synthesis of these studies.

The third identified approach was a comprehension based instruction that was used in two studies to teach science textbook comprehension to students with DD. Students in both studies made progress in their target skills and reported positive experiences. Students indicated that the interventions helped them acquire new skills and the teachers reported that the interventions were effective, feasible to implement and that the target skills were socially important. Quality appraisal results indicated that both studies were of high quality. Overall, comprehension based instruction might be an effective method for teaching science text comprehension to students with DD, but they do not currently have an evidence base for their effectiveness for teaching learners *scientific reasoning* (or science inquiry skills). These skills are essential to help learners identify and manipulate variables to identify causal influences, including the ability to generate predictions and the use of evidence to evaluate findings.

Given the small number of studies additional research is needed to establish the effectiveness of comprehension-based instruction in supporting the acquisition and understanding of science vocabulary and key concepts.

Ten studies incorporating systematic instruction methodology focused on a population of learners with ID only, one with ASD only, and nine with ASD and ID. Three studies did not report diagnosis. Two studies using self-directed learning focused on students with ID only, one with ASD only, and two with ASD and ID. Both studies focusing on comprehension based instruction recruited students with ASD only. There does not appear to be a pattern in the use of specific teaching procedures using self-directed teaching or systematic instruction on their effectiveness dependent on diagnosis (see Tables 1, 2, S4 and S5). Interventions were successfully implemented with students with ASD, ID, and ASD and ID. Only students with ASD were included in studies using comprehension-based instruction (see Table 3 and S6) and so these approaches need to be examined with children with other labels. There was limited availability of information on participants' cultural and ethnic/'racial' origin in the studies included in the present systematic review. Thus, the applicability of findings across diverse groups is unknown.

Implications for teaching science to students with DD

While the dominant perspective in the field of mainstream science education is heavily influenced by teaching methodologies based on a cognitive approach (McGinnis & Kahn, 2014), the majority of studies reported in the present systematic review are consistent with the behavioral approach. Very few studies reported findings from teaching programmes designed from a more constructivist perspective. This might be related to the nature of students with disabilities and their learning, but is more likely a direct reflection of the preference of researchers in special education to favour behavioral approaches as their preferred theoretical framework. The dominant view in the special education field, therefore, is that explicit/systematic instruction is the most effective approach to teaching a range of new skills to students with disabilities (Spooner et al., 2017). The current review suggests teaching methods based on behavioral approaches are likely to be effective strategies for teaching science skills and knowledge to students with DDs.

At this point, it is also important to identify a further limitation of the existing evidence base and thus a further note of caution. A majority of the research on science and DD has emanated from the same extended research group in the USA. This body of work and the commitment of the researchers is commendable, but there is then a need for extensive replication and for more researchers in special education to research science, and a need for more science educators to research science learning and teaching for students with DD.

The opening to this paper provided an overview of the aims of science education and the two main approaches to teaching science, including a review of the features of inquiry-based teaching, the most common approach promoted by science educators and policymakers. The main aim of science education is to enable students to understand some of the ‘big ideas’ in science, and to equip students with the necessary inquiry skills to enable them to work scientifically to answer questions and understand the natural world. These principles apply equally to students with DD. A distinction was also made between the pedagogy of science education and the epistemology of science as a discipline (i.e. a distinction between how pupils *learn* about science compared to how pupils are able to put their learning into practice by *working* scientifically [Kirschner et al., 2006]). Many science educators believe that students learn science most effectively through first-hand practical experiences of carrying out scientific inquiry work (i.e. pupils *learn* science by *doing* science), and this has become the accepted strategy with science researchers and eductaors. However, despite its widespread acceptance, there is no convincing research evidence to support the superiority of inquiry-based teaching strategies compared to more direct (systematic) instructional approaches (Kirschner et al., 2006; Novak, 1988; Mayer, 2004). Evidence from the current review, together with findings from trails in mainstream school settings (Cobern et al., 2010), indicates that systematic and self-directed (inquiry) modes of instruction can be effective approaches for teaching science to students with DD, and that these students are likely to be able to carry out science inquiry work with some degree of independence.

Interestingly, some of the systematic instruction and self-directed (inquiry-based) programmes identified during this review (for example, Jiminez et al., 2014) show positive outcomes with respect to teaching students science knowledge and inquiry skills. Teaching strategies such as these are likely to be promising approaches to teaching science to students with DD, including teaching relevant knowledge and inquiry skills to enable learners to *work* scientifically to help them answer testable questions and gather information. It is important to note, however, that students with disabilities generally require additional support to access inquiry-based instructional tasks and that their science gains increase when components of explicit instruction are used (Rizzo & Taylor, 2016).

Evidence from our review indicates that comprehension-based instruction may be an effective teaching strategy to help students understand science texts. However, none of these comprehension-based studies focused on teaching science inquiry skills to learners. They cannot, therefore, stand alone as instructional strategies and meet the aim of improving the provision for teaching scientific content and skills through more practical scientific experiences without additional provision for teaching inquiry skills. Although this is certainly a practical proposition for science teachers, the utility of combining two methods of instruction to meet one educational goal is low. Systematic instruction and self-directed inquiry may offer a more efficient way forward for teachers.

More recent research from cognitive and developmental psychology has identified a set of core skills in initial science learning that highlights the importance of students' language and observational skills in developing conceptual and procedural understanding (Tolmie et al., 2016). This focus on core skills recognizes the need for the systematic introduction of scientific language to students alongside observations and practical tasks, especially for very young children. The provision of graded tasks, featuring the teaching of specific language and observational tasks, is an important feature of some systematic instructional programmes, and it is reasonable to propose, therefore, that teaching approaches based on systematic instruction will support these emergent core skills in science.

Implications for future research

There is limited evidence for the effectiveness of inquiry-based approaches in the literature. This might be due to the difficulty of implementing this type of teaching approach with students with DD, and/or they cannot be successfully operationalized for this population. The research literature in science for students with DD has been dominated by researchers working from a behavioral tradition. More research is now needed to examine the potential of using inquiry-based science teaching for students with DD, including gathering information on teachers' attitudes, practical implications, and social validity.

More research is also needed on the impact of comprehensive science curricula for students with moderate and severe DD throughout primary and secondary education. Two studies included in the current review evaluated the effectiveness of a systematic instruction science curriculum for students with DD (Jimenez et al., 2014; Smith et al., 2013b). More research (including RCTs) is needed to establish the effectiveness of these programmes, including the ability of students to generalize inquiry skills across different science topics. Some of the approaches in this review focus on developing basic science inquiry skills (e.g.

simple predictions, observations, measuring and recording skills) across a range of investigational work (e.g. exploration, classifying and fair tests inquiry tasks [Goldsworthy et al., 2000]). More research is needed to assess the provision for a wider range of science skills and types of investigation contained within science teaching programmes for learners with DD.

Due to the extensiveness of science content, some relevant articles may not have been identified during database searches. This is especially true for studies targeting a variety of educational targets where only one or two participants were working on science related content. Eight studies included in the review were identified from reference lists instead of via the original searches. For example, in Jameson, McDonnell, Johnson, Risen and Polychronis (2007) the word 'science' (or any other related search term) was not used in the title, abstract or keywords and so it was not recognized during database searches. Although it is possible that some similar studies will have been missed, the systematic review method was designed to identify studies using a range of processes to reduce the risk of omission.

References

*Agran, M., Cavin, M., Wehmeyer, M., & Palmer, S. (2006). Participation of students with Moderate to severe disabilities in the general curriculum: The effects of the self-determined learning model of instruction. Research & Practice for Persons with Severe Disabilities, 31(3), 230-241.

American Psychiatric Association (2013). Diagnostic and Statistical Manual of Mental Disorders (5th ed.). Washington, DC: American Psychiatric Publishing.

Browder, D. M. (2001). Curriculum and Assessment for Students with Moderate and Severe Disabilities. New York: Guildford Press.

*Browder, D. M., Trela, K., Courtade, G. R., Jimenez, B. A., Knight, V., & Flowers, C. (2010). Teaching mathematics and science standards to students with moderate and severe developmental disabilities. The Journal of Special Education, 20(10), 1-10. DOI: 10.1177/0022466910369942.

Browder, D. M., & Spooner, F. (2011). Teaching Students with Moderate and Severe Disabilities. New York, NY: The Guilford Press.

*Carnahan, C. R., & Williamson, P. S. (2013). Does compare-contrast text structure help students with autism spectrum disorder comprehend science text? Council for Exceptional Children, 79(3), 347-363.

*Carnahan, C. R., Williamson, P., Birri, N., Swoboda, C., & Snyder, K. K. (2016). Increasing comprehension of expository science text for students with autism spectrum disorder. Focus on Autism and Other Developmental Disabilities, 31(3), 208-220. DOI: 10.1177/1088357615610539.

Cobern, W., Schuster, D., Adams, B., Appelgate, B., Skjold, B., Undrieu, A., Loving, C. C., & Gobert, J. D. (2010). Experimental comparison of inquiry and direct instruction in science. Research in Science & Technological Education, 28(1), 81-96.

*Collins, B. C., Evans, A., Creech-Galloway, C., Karl, J., & Miller, A. (2007). Comparison of the acquisition and maintenance of teaching functional and core content sight words in special and general education settings. Focus on Autism and Other Developmental Disabilities, 22(4), 220-235.

*Collins, B. C., Hager, K. L., & Creech-Galloway, C. (2011). Addition of functional content during core content instruction with students with moderate disabilities. Education and Training in Autism and Developmental Disabilities, 46(1), 22-39.

*Collins, B. C., Terrell, M., & Test, D. W. (2017). Using a simultaneous prompting procedure to embed core content when teaching a potential employment skill. Career Development and Transition for Exceptional Individuals, 40(1), 36-44. DOI: 10.1177/2165143416680347.

Courtade, G. R., Spooner, F., & Browder, D. M. (2007). Review of studies with students with significant cognitive disabilities which link to science standards. Research & Practice for Persons with Severe Disabilities, 32(1), 43-49.

*Courtade, G. R., Browder, D. M., Spooner, F., & DiBiase, W. (2010). Training teachers to use an inquiry-based task analysis to teach science to students with moderate and severe disabilities. Education and Training in Autism and Developmental Disabilities, 45(3), 378-399.

Critical Appraisal Skills Programme. (2017). CASP (Randomised Controlled Trial) Checklist. Retrieved from http://www.casp-uk.net/casp-tools-checklists

Department for Education. (2014a, May). The Equality Act and Schools. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/315587/Equality_Act_Advice_Final.pdf

Department for Education. (2014b, September). Children with Special Educational Needs 2014: An Analysis. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/350129/SFR31_2014.pdf

Department for Education. (2014c, December). The National Curriculum in England: Framework Document. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/381344/Master_final_national_curriculum_28_Nov.pd

Department for Education. (2015, January). Special Educational Needs and Disability Code of Practice: 0 to 25 Years. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/398815/SEND_Code_of_Practice_January_2015.pdf

Department for Education. (2017, January). Special Educational Needs in England. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/633031/SFR37_2017_Main_Text.pdf

Every Student Succeeds Act, Pub.L. No. 114-95, § 1177 Stat. (2015).

*Fetko, E. E., Collins, B. C., Hager, K. D., & Spriggs, A. D. (2013). Embedding science facts in leisure skill instruction conducted by peer tutors. Education and Training in Autism and Developmental Disabilities, 48(3), 400-411.

Goldsworthy, A., Watson, R., & Wood-Robinson, V. (2000) Investigations: Developing understanding. Hatfield, UK: The Association for Science Education.

Harlen, W. (2015). Working with Big Ideas of Science Education. Trieste: Global Network of Science Academies (IAP) Science Education Programme.

*Heinrich, S., Collins, B. C., Knight, V., & Spriggs, A. D. (2016). Embedded simultaneous prompting procedure to teach STEM content to high school students with moderate disabilities in an inclusive setting. Education and Training in Autism and Developmental Disabilities, 51(1), 41-54.

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. Council for Exceptional Children, 71(2), 165-179.

*Hudson, M. E., Browder, D. M., & Jimenez, B. A. (2014). Effects of a peer-delivered system of least prompts intervention and adapted science read-alouds on listening comprehension for participants with moderate intellectual disability. Education and Training in Autism and Developmental Disabilities, 49(1), 60-77.

*Jameson, J. M., McDonnell, J., Johnson, J. W., Riesen, T., & Polychronis, S. (2007). A comparison of one-to-one embedded instruction in the general education classroom and one-to-one massed practice instruction in the special education classroom. Education and Treatment of Children, 30(1), 23-44.

*Jimenez, B. A., Browder, D. M., & Courtade, G. R. (2009). An exploratory study of self-directed science concept learning by students with moderate intellectual disability. Research & Practice for Persons with Severe Disabilities, 34(2), 33-46.

*Jimenez, B. A., Browder, D. M., Spooner, F., & DiBiase, W. (2012). Inclusive inquiry science using peer-mediated embedded instruction for students with moderate intellectual disability. Council for Exceptional Children, 78(3), 301-317.

*Jimenez, B. A., Lo, Y., & Saunders, A. (2014). The additive effects of scripted lessons plus guided notes on science quiz scores of students with intellectual disabilities and autism. Journal of Special Education, 47(4), 231-244. DOI: 10.1177/0022466912437937.

*Johnson, J. W., McDonnell, J., Holzwarth, V. N. & Hunter, K. (2004). The efficacy of embedded instruction for students with developmental disabilities enrolled in general education classes. Journal of Positive Behavior Interventions, 6(4), 214-227.

*Karl, J., Collins, B. C., Hager, K. D., & Jones Ault, M. (2013). Teaching core content embedded in a functional activity to students with moderate intellectual disability using a simultaneous prompting procedure. Education and Training in Autism and Developmental Disabilities, 48(3), 363-378.

Kirschner, P. A., Sweller, J., & Clarke, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential and inquiry-based teaching. Educational Psychoogist, 41(1), 75-86.

*Knight, V. F., Smith, B. R., Spooner, F., & Browder, D. (2012). Using explicit instruction to teach science descriptors to students with autism spectrum disorder. Journal of Autism and Developmental Disorders. 42(3), 378-389. DOI: 10.1007/s10803-011-1258-1.

*Knight, V. F., Spooner, F., Browder, D. M., Smith, B. R. & Wood, C. L. (2013). Using systematic instruction and graphic organizers to teach science concepts to students with autism spectrum disorders and intellectual disability. Focus on Autism and Other Developmental Disabilities, 28(2), 115-126. DOI: 10.1 177/108835761247530 I.

*Knight, V. F., Wood, C. L., Spooner, F., Browder, D. M., & O’Brien, C. P. (2014). An exploratory study using science eTexts with students with autism spectrum disorder. Focus on Autism and Other Developmental Disorders, 1-14. DOI: 10.1177/1088357614559214.

*Knight, V. F., Creech-Galloway, C. E., Karl, J. M., & Collins, B. C. (2017). Evaluating supported eText to teach science to high school students with moderate intellectual disability. Focus on Autism and Other Developmental Disabilities, 1-10. DOI: 10.1177/1088357617696273.

Mayer, R. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. American Psychologist, 59, 14-19.

*McDonnell, J., Johnson, J. W., Polychronis, S., Riesen, T., Jameson, M., & Kercher, K. (2006). Comparison of one-to-one embedded instruction in general education classes with small group instruction in special education classes. Education and Training in Developmental Disabilities, 41(2), 125-138.

McGinnis, J. R., & Kahn, S. (2014). Special needs and talents in science learning. In. Lederman, N. G., & Abell, S. K. (Eds.), Handbook of research on science education, volume II (pp. 223-245). New York: Routledge.

*McMahon, D. D., Cihak, D. F., Wright, R. E., & Bell, S. M. (2016). Augmented reality for teaching science vocabulary to postsecondary education students with intellectual disabilities and autism. Journal of Research on Technology in Education, 48(1), 1-19. DOI: 10.1080/15391523.2015.1103149.

Mechling, L. C., Gast, D. L., & Fields, E. A. (2008). Evaluation of a portable DVD player and system of least prompts to self-prompt cooking task completion by young adults with moderate intellectual disabilities. The Journal of Special Education, 42(3), 179-190. DOI: 10.1177/0022466907313348.

*Miller, B., & Taber-Doughty, T. (2014). Self-monitoring checklists for inquiry problem-solving: functional problem-solving methods for students with intellectual disability. Education and Training in Autism and Developmental Disabilities, 49(4), 555-567.

*Miller, B., Doughty, T., & Krockover, G. (2015). Using science inquiry methods to promote self-determination and problem-solving skills for students with moderate intellectual disability. Education and Training in Autism and Developmental Disabilities, 50(3), 356-368.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. The PRISMA Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med, 6(7), 1-6. DOI: 10.1371/journal.pmed.1000097.

National Research Council. (1996). National Science Education Standards. Washington, DC: The National Academies Press.

National Research Council. (2013). Next Generation Science Standards: For States, By States. Washington, DC: The National Academies Press.

No Child Left Behind Act of 2001. Pub. L. No.' 107-110, 115 Stat. 1425 (2002).

Novak, J. D. (1988). Learning science and the science of learning. Studies in Science Education, 15, 77-101.

*Riesen, T., McDonnell, J., Johnson, J. W., Polychronis, S., & Jameson, M. (2003). A comparison of constant time delay and simultaneous prompting within embedded instruction in general education classed with students with moderate to severe disabilities. Journal of Behavioral Education, 12(4), 241-259.

*Riggs, L., Collins, B. C., Kleinert, H., & Knight, V. F. (2013). Teaching principles of heredity to high school students with moderate and severe disabilities. Research & Practice for Persons with Severe Disabilities, 38(1), 30-43.

Rizzo, K. L., & Taylor, J. C. (2016). Effects of inquiry-based instruction on science achievement for students with disabilities: an analysis of the literature. Journal of Science Education for Students with Disabilities, 19(1), 1-16.

*Roberts, V., & Joiner, R. (2007). Investigating the efficacy of concept mapping with pupils with autistic spectrum disorder. British Journal of Special Education, 34(3), 127-135.

Sawyer, R. K. (2008). Optimising learning: implications of learning sciences research. In Innovating to Learn, Learning to Innovate, OECD Publishing, Paris, https://doi.org/10.1787/9789264047983-4-en.

Science Council. (n.d.). Our definition of science. Retrieved from http://sciencecouncil.org/about-science/our-definition-of-science/

*Smith, B. R., Spooner, F., & Wood, C. L. (2013a). Using embedded computer-assisted explicit instruction to teach science to students with autism spectrum disorder. Research in Autism Spectrum Disorder, 7(3), 433-443. DOI: 10.1016/j.rasd.2012.10.010.

*Smith, B. R., Spooner, F., Jimenez, B. A., & Browder, D. (2013b). Using an early science curriculum to teach science vocabulary and concepts to students with severe developmental disabilities. Education and Treatment of Children, 36(1), 1-31.

Snyder, T. D., de Brey, C., & Dillow, S. A. (2018). Digest of Education Statistics, 2016 (NCES 2017-094). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Spooner, F., Knight, V., Browder, D., Jimenez, B., & DiBiase, W. (2011). Evaluating evidence-based practice in teaching science content to students with severe developmental disabilities. Research & Practice for Persons with Severe Disabilities, 36(1-2), 62-75.

Spooner, F., McKissick, B. R., Knight, V., & Walker, R. (2014). Teaching Science Concepts. In Browder, D., & Spooner, F. (Ed.), More Language Arts, Math, and Science for Students with Severe Disabilities. (pp. 215-236). Baltimore, MD: Paul H. Brookes Publishing Co.

Spooner, F., & Browder, D. M. (2015). Raising the bar: significant advances and future needs for promoting learning for students with severe disabilities. Remedial and Special Education, 36(1), 28-32.

Spooner, F., McKissick, B. R., & Knight, V. F. (2017). Establishing the state of affairs for evidence-based practices in students with severe disabilities. Research and Practice for Persons with Severe Disabilities, 42(1), 8-18.

Steffe, L., & Gale, J. (eds) (1995). Constructivism in education. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Taber, T. A., Alberto, P. A., Hughes, M., & Seltzer, A. (2002). A strategy for students with moderate disabilities when lost in the community. Research & Practice for Persons with Severe Disabilities, 27(2), 141-152.

The Nation's Report Card. (n.d.). 2015 Science Assessment. Retrieved from https://www.nationsreportcard.gov/science_2015/#?grade=4

Tolmie, A. K., Ghazali, Z. & Morris, S. (2016). Children's science learning: A core skills approach. British Journal of Educational Psychology, 86, 481-497

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2017). A Guide for Ensuring Inclusion and Equity in Education. Paris, France: UNESCO.

Supplementary information liked to the online version of the paper at Wiley-Blackwell:

- Table S1
- Table S2
- Table S3
- Table S4
- Table S5
- Table S6

Table 1
Summary table of studies using systematic instruction procedures.

Source and Origin	Participants	Target skills	Intervention	Design	Science outcomes	Students' and teachers' experiences	Quality appraisal rating
Browder et al. 2010 USA	- 21 students (11 with ASD and 10 with moderate to severe ID) - 12 males and 9 females - 14-21 years - 7 Caucasian, 1 Hispanic and 13 African American - English was a primary language for all participants - IQ: 33-53 (mean 42.90)	Inquiry skills (task analysis of steps to participate in the inquiry lesson on magnetism) and science vocabulary	Task analyzed inquiry based instruction	Quasi experimental design	Baseline Students scored mean 41.9% (SD 11.5) of correct answers for the science test with mean 56.5% (SD 15.3) for inquiry subscale and mean 38.3% (SD 13.2) for science vocabulary subscale. Outcomes At post-test students scored mean 57.6% (SD 22.1) at the science test with mean 70.5% (SD 21.7) for inquiry subscale and mean 54.4% (SD 23.0) for science vocabulary subscale. Overall, students showed 15.7% gain at post-test compared to baseline at the science assessment - 14% gain at the inquiry subtest and 16.1% gain at science vocabulary subscale. Maintenance and generalization Not assessed	Results reported in relation to mathematics and science targets. Teachers' perceptions of the training and interventions were assessed using a survey with a rating scale. They agreed that both interventions (maths and science) were beneficial for their students and practical to implement. Teachers indicated that the materials were helpful, and time spent on practice with the researcher was useful.	8/9 (CASP form for randomized controlled trials)
Collins et al. 2007 USA	- Targets for 1 student were science related. - 1 male student with ID - 9 years - No ethnic/'racial' background or primary language reported	Functional and core science vocabulary/ sight words	Compared three interventions: 1. simultaneous prompting with massed trial instruction in a resource room 2. simultaneous prompting with	Adapted alternating treatments design replicated across three instructional conditions and four participants (although only one had science related targets)	Baseline Student's rate of responding was 0%. Outcomes The student reached mastery criterion only for one word set (functional content) in the embedded instruction condition. Maintenance and generalization The student maintained acquired knowledge for functional content in the embedded instruction condition with 100% accuracy only for six (17%) out of 35 maintenance sessions.	Not reported	Overall 20/21 Science targets only 19/21 (Horner et al., 2005)

Source and Origin	Participants	Target skills	Intervention	Design	Science outcomes	Students' and teachers' experiences	Quality appraisal rating
	- IQ: 50		distributed trial instruction in general education setting 3. embedded instruction in a general education classroom				
Collins et al. 2011 USA	- 3 students (1 with ASD and 1 with Down Syndrome (DS); no diagnosis reported for the third student) - 2 males and one female - 14-15 years - No ethnic/'racial' background or primary language reported - IQ: 41-55 (mean 47.67)	<u>Chemical and physical properties</u> of elements in the Periodic Table (gases, liquids and solids)	Constant time delay procedure	Multiple probe design across behaviors (language arts, science and math) replicated across participants	<u>Baseline</u> Students had between 11.1% and 75% (mean 35.6%) of correct responses for core content and between 33% and 62.9% (mean 51.4%) for functional content. <u>Outcomes</u> After the intervention was implemented students met mastery criterion in four to 69 sessions (mean 28.3) for core content and in four to 32 sessions (mean 13.3) for functional content. <u>Maintenance and generalization</u> Students maintained core content with 44.4% to 100% accuracy (mean 80.8%) and functional with 33.3% to 100% accuracy (mean 77.8%). Students' scores during generalization increased from mean 20.1% to 88.9%.	Not reported	<u>Overall</u> 20/21 <u>Science targets only</u> 20/21 (Horner et al., 2005)
Collins et al. 2017 USA	- 4 students with ID - 2 females and 2 males - 16-19 years	<u>Science concepts</u> (Photosynth-esis core content) embedded in	Simultaneous prompting procedure used to embed core content in teaching	Multiple probe across participants design with pre- and post-test measures	<u>Baseline</u> Students answered between one and two (out of six) questions correctly (mean 1.3). <u>Outcomes</u> Responding improved at post-test compared to baseline for all students, but none of the students reached mastery criterion. Students answered between four and five (out of	<u>Results reported in relation to science targets and practical skill training</u>. Students' experiences of the intervention were assessed using a questionnaire. Students reported that they enjoyed the intervention and that	<u>Overall</u> 21/21 <u>Science targets only</u> 21/21

Source and Origin	Participants	Target skills	Intervention	Design	Science outcomes	Students' and teachers' experiences	Quality appraisal rating
	- 3 Hispanic and 1 African American - No primary language reported - GIA: 62-71 (mean 67)	task analysis for plant care	practical skill (plant care)	of the non-target information (science concepts)	six) questions correctly (mean 4.5). Since none of the students reached mastery criterion, students had additional simultaneous prompting procedure training for core content only and reached the mastery criterion within three sessions. <u>Maintenance and generalization</u> All participants maintained acquired knowledge with 100% accuracy over time.	they learned about photosynthesis. Three students indicated that they would use acquired skills in the future and one said they would not.	(Horner et al., 2005)
Courtade et al. 2010 USA	- 8 students with ID - 4 females and 4 males - 11-15 years - 5 African American, 2 Caucasian and 1 Hispanic - English was a primary language for all participants - IQ: 39-54 (mean 44.14 - not reported for one student)	<u>Inquiry skills</u> (task analysis of steps to participate in inquiry lessons) and <u>science vocabulary</u>	Multi-component training in task analyzed inquiry-based instruction for teachers, including: fidelity checklist, training manual, verbal explanation of content, video modelling and feedback from the researchers	Multiple probe across participants single subject design	<u>Baseline</u> Students' scores were between one and three correct (out of 12). <u>Outcomes</u> After the intervention was implemented students' scores ranged between three and 12, with majority of scores being nine (75%) or higher. <u>Maintenance and generalization</u> Maintenance probes were conducted with only two students. Their mean score was 10 (range 9-11). One of the teachers reported that her student used new science terms in a context different to the science lesson.	Parents and teachers views, and experiences were assessed using surveys with a rating scale and open-ended questions. <u>The parents</u> agreed that it is important for their children to learn science and that they should have science lessons every day. Parents also agreed that it is important that science instruction is recommended by the National Science Education Standards. Parents reported that their children showed interest in science skills. <u>Teachers</u> responding on the validity survey was in the range of 5-6 (6-point rating scale). Teachers responding on the feasibility survey was in the range of 3-5 (5-point rating scale).	<u>Overall</u> 21/21 <u>Science targets only</u> 21/21 (Horner et al., 2005)
Fetko et al. 2013 USA	- 3 students (2 with ID and 1 with ASD) - 2 males and 1 female - 12-14 years - No ethnic/'racial' background or	<u>Science vocabulary</u>	Simultaneous prompting procedure with core content (science vocabulary) embedded as non-target	Multiple probe design across participants with pre- and post-test measures of the non-target information	<u>Baseline</u> All three students scored 0% during the baseline probe. <u>Outcomes</u> Two students reached 100% during the post-test probe and one student scored 0%. <u>Maintenance and generalization</u> Not assessed.	Not reported	<u>Overall</u> 19/21 <u>Science targets only</u> 16/21 (Horner et al., 2005)

Source and Origin	Participants	Target skills	Intervention	Design	Science outcomes	Students' and teachers' experiences	Quality appraisal rating
	primary language reported - IQ scores not reported		information while teaching a leisure skill activity (UNO game)	(science vocabulary)			
Heinrich et al. 2016 USA	- Targets for 1 student were science related - 1 male student with moderate ID - 17 years - No ethnic/'racial' background or primary language reported - IQ: 53	Science vocabulary and Punnett Square	Embedded simultaneous prompting procedure	Multiple probe across participants design with concurrent demonstration across two skills per student	Baseline The student scored correctly to 0% of probes for both discrete and chained tasks. Outcomes The student reached mastery criterion for science vocabulary (discrete task) in seven sessions and for Punnett Square (chained task) in five sessions. Maintenance and generalization The student maintained acquired content with 100% accuracy after a month and generalized some content to other contexts. He also showed some generalization of acquired skills during the state assessment, scoring mean of 60% of correct responses for discrete tasks and 100% for chained tasks.	Data reported in relation to all participants in the study. Peers' and general education teacher's attitudes towards students with disabilities were assessed using a survey. Before the intervention out of 17 peers, 12 said that students with ID should attend general education classrooms. After the intervention, the number increased to 15, all peers also indicated that students with disabilities can learn core content. Before the intervention sixteen students thought that students with disabilities should be taught core content and after the intervention all seventeen students said that they should. The number of peers agreeing with the following benefits of inclusion also increased after the intervention was implemented: social interactions, academic skills acquisition, communication skills and self-esteem. The general education teacher indicated that she thought that students with disabilities should attend general education classes, can learn core content and would benefit from the inclusion. Following the intervention, she also indicated that students with disabilities can learn core content at a modified pace.	Overall 20/21 Science targets only 18/21 (Horner et al., 2005)

Source and Origin	Participants	Target skills	Intervention	Design	Science outcomes	Students' and teachers' experiences	Quality appraisal rating
Hudson et al. 2014 USA	- 3 students with ID - 2 females and 1 male - No age range and no IQ scores reported - No ethnic/'racial' background or primary language reported	Listening comprehens-ion of science content	Peer-delivered system of least prompts with adopted science read-alouds	Multiple probe design across participants	Baseline Students responded correctly to 18-27% of questions correctly (mean 22.7%). Outcomes After the intervention was implemented, students' rate of correct responses increased to 63-79% (mean 71.3%). Maintenance and generalization Rate of responding during generalization probes did not exceed baseline levels for all three students.	Attitude surveys Peer tutors at pre-test indicated that they had limited contact with people with disabilities and the majority were not sure if they would talk to a student with a disability. At post-test the majority of peer tutors indicated that they would talk and eat lunch with a student with a disability. Social validity forms (with a rating scale) Teachers either agreed or strongly agreed that students with disabilities can learn in general education classes, that the peer-delivered instruction is effective in teaching new content to students with disabilities and that they would use and recommend the intervention. Peer tutors reported that they enjoyed their role, would like to do it again in the future and would recommend the intervention. One peer tutor said that the intervention required a lot of work while the other said it did not.	21/21 (Horner et al., 2005)
Jameson et al. 2007 USA	- Targets for 1 student were science related - 1 male student with DS - 15 years - Caucasian - No primary language reported - IQ: 46	Science vocabulary on states of matter content	Comparison of two interventions: 1. One-to-one embedded instruction 2. One-to-one massed trials instructional format	Single subject alternating treatment design	Baseline The student responded correctly to 0% of probes Outcomes Both interventions were effective in teaching science vocabulary to the participant. The student reached mastery criterion in fewer sessions in the one-to-one embedded instruction condition - 255 trials (around 19 sessions) than in the one-to-one massed trials instruction condition - 342 trials (around 27 sessions). Maintenance and generalization Not assessed.	Results reported in relation to all students. Teachers' and paraprofessionals' perceptions of the intervention were assessed using a questionnaire with a rating scale. They reported that the embedded instruction was effective and practical. Teachers and paraprofessionals also indicated that the prompting procedure was feasible, useful for the students and helped them with inclusion in general education classrooms.	Overall 21/21 Science targets only 20/21 (Horner et al., 2005)

Source and Origin	Participants	Target skills	Intervention	Design	Science outcomes	Students' and teachers' experiences	Quality appraisal rating
Jimenez et al. 2009 USA	- 3 students with ID - 2 females and 1 male - 11-13 years - No ethnic/'racial' background or primary language reported - IQ: 48-54 (mean 51.3)	Self-directed inquiry (task analysis on using a KWHL chart - What we know?; What we want to know?; How to find out?; What was learned?) and science concepts	Multicomponent training package (multiple exemplar training, time delay and KWHL chart)	Multiple probe design across two science concepts with a concurrent between participant replication	Baseline Students were not correct with any steps of the task analysis for both concepts. Outcomes After the intervention was implemented students reached mastery criterion for the first concept in one to five sessions (mean 3.3). Two students exhibited spontaneous generalization across concepts and reached mastery criterion for the second concept before intervention was implemented. The third student reached mastery criterion for the second concept within one session. Maintenance and generalization During maintenance probes students responded correctly to all probes. Students generalized acquired knowledge across materials and the second concept. They also generalized the use of KWHL chart to the general education classroom.	Students' and teachers' views were assessed using the adopted intervention rating profile with a rating scale. The teachers strongly agreed to all statements about intervention's acceptability, procedures and outcomes. The students indicated that they enjoyed the intervention to learn science and liked using KWHL charts. The students reported that the intervention might also be beneficial for other students.	Overall 21/21 (Horner et al., 2005)
Jimenez et al. 2012 USA	- 5 students with ID - 2 females and 3 males - 11-14 years - No ethnic/'racial' background or primary language reported - IQ: 34-55 (mean 46.2)	Science vocabulary and concepts and the use of KWHL chart during inquiry lessons	Peer-mediated embedded instruction with time delay	Multiple probe across three science units with between participant replications	Baseline Students had between one and six correct responses (mean 2.6) for Unit 1, between zero and six (mean 2.3) for Unit 2 and between two and seven (mean 3.4) for Unit 3. Outcomes After the intervention was implemented students had between three and eight correct responses (mean 7.2) for Unit 1, between two and eight (mean 6.4) for Unit 2 and between four and eight (mean 6.57) for Unit 3. Maintenance and generalization Data not clearly reported.	There was an increase in surveys' scores (5-point rating scale) from pre- to post-test. Peer tutors' scores increased from 3.2 to 4.6 and students' scores increased from 3.5 to 4.7. During the focus group peer tutors indicated that they enjoyed the intervention and wanted to continue with it. Peer tutors also indicated that the intervention was beneficial to them. In the feasibility survey the teachers agreed that the intervention was socially important, effective and practical to implement. Grades of the peer tutors remained the same throughout the intervention.	Overall 21/21 (Horner et al., 2005)
Jimenez et al. 2014 USA	- 3 students with ASD and ID	Three science content units	Scripted lessons and scripted	Multiple probe across science content units	Baseline Students had between zero and seven correct responses (mean 2.6) for Unit 1,	Teachers' views and experiences of the intervention were assessed using social validity questionnaires. They	Overall 20/21

Source and Origin	Participants	Target skills	Intervention	Design	Science outcomes	Students' and teachers' experiences	Quality appraisal rating
	- 2 males and 1 female - 9 years - African American - No primary language reported - IQ: 71-99 (mean 86.67)	including inquiry skills, science concepts and vocabulary	lessons with guided notes	design with replication across students	between zero and 10 (mean 4.4) for Unit 2 and between zero and 10 (mean 3.9) for Unit 3. <u>Outcomes</u> When scripted lessons were introduced students' rate of correct responses improved to between zero to 10 (mean 5.8) for Unit 1, between one and 10 (mean 6.4) for Unit 2 and between one and 10 (mean 7.5) for Unit 3. Once scripted lessons with guided notes were introduced, students had between zero and 10 correct responses (mean 5.9) for Unit 1, between two and 10 (mean 7.6) for Unit 2 and between three and 10 (mean 8.9) for Unit 3. <u>Maintenance and generalization</u> Students maintained their rate of responding over time apart from Student 3 for one of the units.	reported that both interventions were effective in teaching science to the students but that the scripted lesson condition was preferred. Scripted lessons with guided notes were reported to be more time consuming.	(Horner et al., 2005)
Johnson et al. 2004 USA	- Targets for 1 student were science related. - Female with DD (exact diagnosis not reported) - 9 years - No ethnic/'racial' background or primary language reported - IQ: 59	<u>Science concepts</u>	Embedded instruction implemented in general education classroom (constant time delay, error correction and reinforcement)	Multiple baseline across behaviors design	<u>Baseline</u> The student had 0% correct responses at baseline probes for all three units. <u>Outcomes</u> After the intervention was implemented the student reached the mastery criterion for all three units in 4 – 7 sessions. <u>Maintenance and generalization</u> Student's rate of responding was maintained for two units (maintenance data for third unit was not collected) over time.	<u>Results reported in relation to all students.</u> Teachers' and paraprofessionals' views and opinions of the intervention were assessed using questionnaires with a rating scale. They reported that the intervention was effective, it met students' needs and it was not very disruptive to the rest of the class. Staff members indicated that they were likely to use the intervention in the future.	<u>Overall</u> 19/21 <u>Science targets only</u> 19/21 (Horner et al., 2005)
Karl et al. 2013 USA	- 4 students with ID - 3 males and 1 female - 15-18 years	<u>Science concepts</u>	Simultaneous prompting procedure used to teach core content within a	Multiple probe design across behaviors replicated across participants	<u>Baseline</u> Students had 0% of correct responses. <u>Outcomes</u> Students reached mastery criterion in four to 23 sessions (mean11.5). <u>Maintenance and generalization</u> Three participants maintained acquired knowledge with 100% accuracy after one, three and five	Not reported	<u>Overall</u> 20/21 <u>Science targets only</u> 20/21

Source and Origin	Participants	Target skills	Intervention	Design	Science outcomes	Students' and teachers' experiences	Quality appraisal rating
	- No ethnic/'racial' background or primary language reported - IQ: 41-55 (mean 48)		functional activity (cooking)		weeks (no data reported for one student) and all student generalized target skills with 100% accuracy to different materials.		(Horner et al., 2005)
Knight et al. 2012 USA	- 3 students with ASD - 3 males - 5-7 years - No ethnic/'racial' background or primary language reported - IQ: 53 and 62 (not reported for 1 student) (mean 57.5)	Science descriptors	Explicit instruction (model-lead-test strategy)	Multiple probe across behaviors with concurrent replication across participants design	Baseline Students correctly responded to between zero and two science descriptors (mean 0.7) for Set 1, between zero and two (mean 1.1) for Set 2 and between zero and three (mean 0.8) for Set 3. Outcomes After the intervention was implemented students reached mastery criterion in 16-22 sessions (mean 18.3) for Set 1, in 10-14 sessions (mean 12.7) for Set 2 and in 12-18 sessions (mean 14.3) for Set 3. Maintenance and generalization Two students maintained high rate of responses over time and all of the students generalized acquired knowledge across different materials.	Students' and teachers' views and experiences were assessed using questionnaires. Students' impressions of the intervention were positive, and they indicated willingness to participate in the future research. The teacher strongly agreed that targets were socially important to the students and that the intervention was a good use of time. She also indicated she would be interested in taking part in future research. The teacher agreed that acquired targets generalized to other inquiry content but not to other settings and she would use explicit instruction in the future.	Overall 21/21 (Horner et al., 2005)
Knight et al. 2013 USA	- 3 students with ID and ASD - 1 female and 2 males - 13-14 years - No ethnic/'racial' background or primary language reported	Science concepts	Treatment package of systematic instruction (constant time delay, examples and non-examples, graphic organizers)	Multiple probe across participants design	Baseline Students had between zero and seven correct responses at the task analysis (mean 2.8). Outcomes After the intervention was implemented students reached mastery criterion in seven to eight sessions (mean 7.7). Maintenance and generalization Data collected only for two students – they maintained high rate of correct responses over time.	Not reported	Overall 21/21 (Horner et al., 2005)

Source and Origin	Participants	Target skills	Intervention	Design	Science outcomes	Students' and teachers' experiences	Quality appraisal rating
	- IQ: 40-55 (mean 46.33)						
Knight et al. 2014 USA	- 4 students with ID and ASD - 1 female and 3 males - 11-14 years - African American - No primary language reported - IQ: 53-67 (mean 59.5)	Science vocabulary and concepts comprehen-sion	Book Builder (BB) - three phases: - BB only - BB and explicit instruction (EI) - BB, EI and referring to definition	Multiple probe across participants with an embedded ABCD design	Baseline Students responded correctly to between 8.3% and 33.3% (mean 20.7%) of vocabulary questions, between 16.7% and 40% (mean 28%) of comprehension questions and between 10% and 50% (mean 29.9%) for application questions. Outcomes After the first phase of the intervention (BB only) was implemented students responded correctly to between 22.2% and 44.5% (mean 33.3%) of vocabulary questions, between 25% and 55.6% (mean 43.8%) of comprehension questions and between 0% and 75% (mean 27.1%) of application questions. After the second phase of the intervention (BB and explicit instruction) was introduced students responded correctly to between 16.7% and 66.7% (41.7%) of vocabulary questions, between 50% and 77.8% (mean 62.5%) of comprehension questions and between 0% and 66.67% (mean 45.8%) for comprehension. After the third phased of the intervention (BB, EI and referring to definition) was implemented students responded correctly to between 16.67% and 100% (mean 64.2%) of vocabulary questions, between 50% and 80% (mean 60%) of comprehension questions and between 40% and 100% (mean 67.5%) of application questions. Maintenance and generalization Students' responding during maintenance probes was between four and seven correct (out of seven). Data not collected for two students.	Students' and teachers' views and experiences were assessed using surveys. The teachers agreed that the intervention was effective, practical and that they would use it in the future. They also agreed that the intervention might be useful for students in other areas. They reported the most helpful resource to be coaches, limited language, summarizing resources and visual cues. One of the teachers reported that the intervention would be more effective if it could respond to students' errors. Students reported to have enjoyed the intervention. They found coaches and hyperlinks to vocabulary to be most helpful.	Overall 20/21 (Horner et al., 2005)

Source and Origin	Participants	Target skills	Intervention	Design	Science outcomes	Students' and teachers' experiences	Quality appraisal rating
Knight et al. 2017 USA	- 4 students with ID - 1 female and 3 males - 18-21 years - White - No primary language reported - IQ: 41-55 (mean 47.5)	Science comprehens-ion skills (vocabulary, comprehens-ion and application probes)	Modified Book Builder (embedded animated coaches, examples and non-examples and referrals to the definitions)	Multiple probe across participants research design	Baseline Students had mean of 1.4 correct responses. Outcomes After the intervention was introduced students met mastery criterion on seven to 11 sessions (mean 9.3). Maintenance and generalization Not assessed.	Not reported	Overall 21/21 (Horner et al., 2005)
McDonnell et al. 2006 USA	- Targets for 2 students were science related. - Both students with DD - Males - 13-15 years - No ethnic/'racial' background or primary language reported - IQ: 50 and 55 (mean 52.5)	Science concepts/ definitions	Comparison of two interventions: 1. One-to-one embedded instruction 2. Small-group spaced-trial instruction	Alternating treatment design	Baseline Students had 0% of correct responses. Outcomes When intervention was implemented students reached mastery criterion for both conditions in 435-585 trials (mean 510). Maintenance and generalization Not assessed.	Not reported	Overall 20/21 Science targets only 20/21 (Horner et al., 2005)
Riesen et al. 2003 USA	- Targets for 2 students were science related - 1 student with ASD and 1 with	Science vocabulary and concepts	Embedded instruction with comparison between constant time delay and	Adapted alternating treatment design	Baseline Students had 0% of correct responses. Outcomes Students reached mastery criterion for simultaneous prompting condition in 17-54 trials (mean 35.5). Due to time constraints only one student reached mastery criterion for constant time delay condition (34 trials). Maintenance and generalization Not assessed.	Not reported	Overall 20/21 Science targets only 20/21

Source and Origin	Participants	Target skills	Intervention	Design	Science outcomes	Students' and teachers' experiences	Quality appraisal rating
	multiple disabilities. - Male and female - 13 years - No ethnic/'racial' background or primary language reported - IQ: 55 and 66 (mean 60.5)		simultaneous prompting				(Horner et al., 2005)
Riggs et al. 2013 USA	- 5 students with moderate to severe disability - 3 males and 2 females - 14-18 years - No ethnic/'racial' background or primary language reported - IQ: 40-76 (mean 50.8)	Science concepts	Constant time delay procedure with examples and non-examples	Multiple probe design replicated across students	Baseline Based on students' responding during baseline probes researchers determined a starting point for all participants. One students started at Level 1, two at Level 2 and two at Level 3. Outcomes Students required between four and 18 sessions to reach mastery criterion (mean 8.6). Maintenance and generalization All students had 100% at 1-week maintenance probes. At 3-week maintenance probes students had between 67% and 100% (mean 93.4%). Students generalized acquired knowledge to novel exemplars.	Not reported	Overall 19/21 (Horner et al., 2005)
Smith et al. 2013a USA	- 3 students (2 with ASD and 1 with ASD and ID) - Males - 11-12 years - 1 Asian/Pacific	Science vocabulary (terms and applications)	Embedded computer-assisted explicit instruction	Multiple probe across participants design	Baseline Students had between one and four (out of 18) correct responses. Outcomes After the intervention was implemented students reached mastery criterion for all three units after six to eight sessions (mean 7). Maintenance and generalization Students had 12-13 correct responses (mean 12.7) at one	Students, peer tutors' and teachers' views and opinions were assessed using questionnaires. Students reported that science is important for all students. They also indicated that the intervention was effective, and they would like to receive more instruction using iPads. The peer	Overall 21/21 (Horner et al., 2005)

Source and Origin	Participants	Target skills	Intervention	Design	Science outcomes	Students' and teachers' experiences	Quality appraisal rating
	Islander, 1 biracial (African American and Caucasian) and 1 Native Hawaiian/Other Pacific - No primary language reported - IQ: 59 and 69 (not reported for 1 student) (mean 64)				week maintenance probe. Their responding decreased by one compared to intervention values during generalization probe.	tutors reported that the intervention was effective, and they would like to use iPads in their own classrooms. They indicated that science education is important for all students and they enjoyed supporting students with disabilities. The teachers reported that the intervention was effective, and it was time well spent. They also expressed their interest of using technology in the classrooms.	
Smith et al. 2013b USA	- 3 students (1 with ID and 2 with multiple disabilities) - 2 females and 1 male - 6-7 years - 1 African American and 2 Caucasian - No primary language reported - IQ scores not reported.	Science concepts	Task analyzed science inquiry lessons	Multiple probe across behaviors with concurrent replication across participants design	Baseline Students responded correctly to between two and three probes (mean 2.3) for Unit 1, between 0.8 and 2.5 (mean 1.6) for Unit 2, between one and 4.2 (mean 2.6) for Unit 3 and between 1.6 and 2.4 (mean 1.9) for Unit 4. Outcomes Students responded correctly to between 4.8 and 6.5 probes (mean 5.9) for Unit 1, between 5.3 and six (mean 5.6) for Unit 2, between 6.2 and 7.4 (mean 6.7) for Unit 3 and between 5.6 and 6.9 (mean 6.1) for Unit 4. Maintenance and generalization Students responding during maintenance probes remained the same or slightly decreased compared to intervention outcomes.	Students' and teachers' views and experiences were assessed using a questionnaire. The students reported that they enjoyed the intervention and would like to do it again in the future. Two students (out of three) said that the intervention was not helpful during other lessons. The teacher strongly agreed that the intervention was a good use of time and she would like to participate in similar projects in the future. The teacher also agreed that targets were important, and she would use some components in the future. She reported that acquired skills didn't generalize to other classes.	Overall 20/21 (Horner et al., 2005)

Table 2

Summary table of studies using self-directed learning procedures.

Source and Origin	Participants	Target skills	Intervention	Design	Science outcomes	Students' and teachers' experiences	Quality appraisal rating
Agran et al. 2006 USA	- Targets for 2 students were science related. - 1 student with ID and 1 with ASD - Male and female - 13-15 years - No ethnic/'racial' background or primary language reported - IQ scores not reported	Inquiry skills for Student 1 and Science concepts for Student 2	Self-determined learning model of instruction - self-monitoring and goal setting	Multiple baseline across individuals design	Baseline Students had between 0% and 25% of correct responses (mean 8.5%). Outcomes After the intervention was introduced students reached mastery criterion in 10-18 sessions (mean 14). Their performance ranged from 13% to 87% (mean 60%) Maintenance and generalization Students maintained acquired skills with between 75% and 87% (mean 82.5%) (one of the students had only one maintenance session).	Students' views and experiences were assessed using self-evaluation forms. One student made no verbal responses to any of the questions. The other student reported that as a result of the intervention: she was working harder in science class, she appreciated having guidelines, she knew what she wanted to know but she indicted that she didn't know what changed with things she did not know before the intervention. Teachers of both students reported improvement in their lesson participation.	Overall 20/21 Science targets only 20/21 (Horner et al., 2005)
McMahon et al. 2016 USA	- 4 students (3 with ID and 1 with ASD) - 1 male and 3 females - 19-25 years - No ethnic/'racial' background or primary language reported - IQ: 48-85 (mean 65.25)	Science vocabulary	Augmented Reality application	Multiple-probe across-behaviors/ Skills design	Baseline Students' average performance for first word list was between 6.7-30%, between 7.5-27.5% for second word list and between 10-20% for the third word list. Outcomes Students reached mastery criterion for the first word list in four to eight sessions (mean 6.5), for the second word list in five to 11 sessions (mean 9.5) and for the third word list in five to 11 sessions (mean 7.5). Maintenance and generalization Not assessed.	Students' views and experiences were assessed using surveys with a rating scale and two open-ended questions. Students reported that the intervention was socially appropriate, helpful, feasible and they would like to use it in the future with other targets. They also reported that hearing the definitions read aloud was easier than reading them and the intervention was enjoyable.	Overall 21/21 (Horner et al., 2005)

Source and Origin	Participants	Target skills	Intervention	Design	Science outcomes	Students' and teachers' experiences	Quality appraisal rating
Miller and Taber-Doughty, 2014 USA	- 3 students with ID - 2 females and 1 male - 12-13 years - 2 Caucasian and 1 Latino - No primary language reported - IQ: 46-64 (mean 55.33)	Inquiry skills (task analyzed)	Self-monitoring checklist and science note-book	Multiple probe design	Baseline Students responded correctly on average to 6.7% steps on the task analysis. Outcomes After the intervention was implemented students' rate of responding improved to 96-100%. Maintenance and generalization Students' responding during generalization probes remained at the same level as during the intervention.	Students' views and experiences were assessed using the social validity interviews revised Treatment Acceptability Rating Form). Students reported that they enjoyed the intervention and wanted to continue and recommend it to others. They reported that the checklist was helpful and that they wanted to use it in the future and that the science notebooks were useful (two students out of three).	Overall 21/21 (Horner et al., 2005)
Miller et al. 2015 USA	- 3 students with ID - 2 females and 1 male - 14-19 years - 2 Caucasian and 1 Latino - No primary language reported - IQ scores not reported	Inquiry skills (task analyzed)	Guided science inquiry and self-monitoring checklist	Multiple probe across participants design	Baseline Students had between 23.3% and 49.53% of steps of the task analysis completed correctly (mean 35.1%). Outcomes After the intervention was implemented students' responding increased to 58.5-95.8% (mean 79.2%). Maintenance and generalization During generalization probes students responding remained high – 77.9-96.9% (mean 89.5).	Students' views and opinions were assessed using questionnaires. Students reported to have enjoyed the intervention and would like to continue. Two students (out of three) indicated that the checklist was helpful and would be useful in other classes, but all students reported they would prefer not to use it.	Overall 21/21 (Horner et al., 2005)
Roberts and Joiner, 2007 UK	- 10 students with ASD - 9 males and 1 female - 11-14 years - No ethnic/'racial' background or primary language reported - IQ: 63-120 (mean 92)	Science concepts and maps	Comparison of two interventions: 1. Concept mapping (experimental) 2. Conventional teaching (control)	Within-participant crossover experimental design	Baseline Students in the concept mapping group scored mean of 29.6 points (SE 7.8) and students in the conventional teaching group scored mean of 47 points (SE 4.2). Outcomes The Wilcoxon signed-rank test determined that the difference in baseline and post-test measures for science questionnaires (concepts) was significantly bigger for experimental (concept mapping) than control (conventional teaching) conditions ($z=2.091$; $p<0.05$; $r=0.66$). This was determined as a large effect size (Cohen's effect size criteria). There was no significant difference between	Not reported	7/8 (CASP form for randomized controlled trials)

Source and Origin	Participants	Target skills	Intervention	Design	Science outcomes	Students' and teachers' experiences	Quality appraisal rating
					experimental and control conditions for concept maps ($z=1.48$; $p>0.05$; $r=0.47$). Maintenance and generalization Not assessed.		

Table 3
Summary of studies using comprehension based instruction.

Source and Origin	**Participants**	**Target skills**	**Intervention**	**Design**	**Science outcomes**	**Students' and teachers' experiences**	**Quality appraisal rating**
Carnahan and Williamson, 2013 USA	- 3 students with ASD - Males - 13 years - No ethnic/'racial' background or primary language reported - No IQ scores were reported.	<u>Science textbook comprehens-ion</u>	Compare-contrast strategy package	Single-subject reversal design	<u>Baseline</u> Students' responding was between 50% and 77% (mean 62.3%). <u>Outcomes</u> After the intervention was implemented students' responding improved to 97%. <u>Maintenance and generalization</u> During maintenance probes students responding remained high at 95-100% (mean 98.3%).	Teacher's views and experiences were assessed using a questionnaire. The Teacher indicated that the intervention targeted important areas for her students, was feasible to implement and increased student's comprehension of science textbooks.	<u>Overall</u> 21/21 (Horner et al., 2005)
Carnahan et al. 2016 USA	- 3 students with ASD - Males - 15-16 years - No ethnic/'racial' background or primary language reported - IQ: 76 (reported only for one student)	<u>Science texts comprehensi-on</u>	Multicomponent text structure intervention (text structure organization and text analysis)	Multiple baseline design	<u>Baseline</u> Students' average rate of responding was 42-54% (mean 49%). <u>Outcomes</u> Students' responding improved to 88-97% (mean 91.7%) when the intervention was implemented. <u>Maintenance and generalization</u> Students' responding during maintenance probes remained high at 95%.	Students' and teachers' views were assessed using questionnaires with a rating scale and open-ended questions. Both students and teachers reported that the intervention was feasible, effective and would likely be continued. Students reported that they would not change the intervention. Teachers felt that the intervention was helpful in learning more about their students' skills.	<u>Overall</u> 21/21 (Horner et al., 2005)

Table S1
Lists of terms used in the database searches.

List 1	List 2
Autis*	Scien*
ASD	Physics
"Autism Spectrum Disorder*"	Chemistry
"Intellectual Disabilit*”	Biology
ID	Plant*
"Mental retardation"	Animal*
"Developmental Disabilit*"	"Human bod*"
"Down syndrome"	Material*
"Pervasive developmental disorder"	Force*
PDD	Earth
Asperger*	Electricity
"Learning Disabilit*"	Acid*
"Learning Difficult*"	Rocks
"Learning Impairment*"	Soil
"Intellectual Deficien*"	Magnet*
"Developmental Impairment*"	Space
Handicap*	Chemical
	Weather
	Season*
	Mass
	Planet*
	"Solar system*"
	"Living organism*"
	Cell*
	Bodypart*
	Fungus
	Insect*
	Temperature
	"Work* scientifically"
	"Scien* enquiry"
	"Scien* inquiry"

	"Scien* Experiment" STEM "Scien* model* and analog* "Scien* pattern-seek*" "Scien* curriculum" "Scien* intervention" "Scien* program*” "Scien* prediction" "Scien* classification" "Scien* test*"

Table S2

Quality appraisal scores of studies incorporating single-case experimental design.

Quality Indicators	Argan et al. 2006	Agran et al. 2006 (science only)	Carnahan and Williamson, 2013	Carnahan et al. 2016	Collins et al. 2007
1. Participants and Setting					
- Participants described	Y	Y	Y	Y	Y
- Selection described	Y	Y	Y	Y	Y
- Setting described	Y	Y	Y	Y	Y
2. Dependent Variable (DV)					
- DV described	Y	Y	Y	Y	N
- Quantifiable index	Y	Y	Y	Y	Y
- DV measurement described	Y	Y	Y	Y	Y
- DV measured repeatedly	Y	Y	Y	Y	Y
- Inter-observer agreement data reported	Y	Y	Y	Y	Y
3. Independent variable (IV)					
- IV described	Y	Y	Y	Y	Y
- IV systematically manipulated	Y	Y	Y	Y	Y
- Procedural fidelity data reported	N	N	Y	Y	Y
4. Baseline					
- DV repeatedly measured prior to IV implementation	Y	Y	Y	Y	Y
- Baseline procedures described	Y	Y	Y	Y	Y
5. Experimental Control					
- Three demonstrations of experimental control	Y	Y	Y	Y	Y
- Design controlled for common threats to internal validity	Y	Y	Y	Y	Y
- Pattern of results demonstrates experimental control	Y	Y	Y	Y	Y
6. External validity					
- Experimental effects replicated across participants, setting, or materials	Y	Y	Y	Y	Y
7. Social validity					
- DV is socially important	Y	Y	Y	Y	Y
- Magnitude of change in the DV from the intervention is socially important	Y	Y	Y	Y	Y
- Implementation of IV is practical and cost effective	Y	Y	Y	Y	Y
- IV implemented over extended time periods, by typical agents, in typical context	Y	Y	Y	Y	Y
Indicators met:	20/21	20/21	21/21	21/21	20/21
Categories met:	6/7	6/7	7/7	6/7	6/7

Quality Indicators	Collins et al. 2007 (science only)	Collins et al. 2011	Collins et al. 2011 (science only)	Collins et al. 2017	Collins et al. 2017 (science only)
1. Participants and Setting					
- Participants described	Y	N	N	Y	Y
- Selection described	Y	Y	Y	Y	Y
- Setting described	Y	Y	Y	Y	Y
2. Dependent Variable (DV)					
- DV described	N	Y	Y	Y	Y
- Quantifiable index	Y	Y	Y	Y	Y
- DV measurement described	Y	Y	Y	Y	Y
- DV measured repeatedly	Y	Y	Y	Y	Y
- Inter-observer agreement data reported	Y	Y	Y	Y	Y
3. Independent variable (IV)					
- IV described	Y	Y	Y	Y	Y
- IV systematically manipulated	Y	Y	Y	Y	Y
- Procedural fidelity data reported	Y	Y	Y	Y	Y
4. Baseline					
- DV repeatedly measured prior to IV implementation	Y	Y	Y	Y	Y
- Baseline procedures described	Y	Y	Y	Y	Y
5. Experimental Control					
- Three demonstrations of experimental control	Y	Y	Y	Y	Y
- Design controlled for common threats to internal validity	Y	Y	Y	Y	Y
- Pattern of results demonstrates experimental control	Y	Y	Y	Y	Y
6. External validity					
- Experimental effects replicated across participants, setting, or materials	N	Y	Y	Y	Y
7. Social validity					
- DV is socially important	Y	Y	Y	Y	Y
- Magnitude of change in the DV from the intervention is socially important	Y	Y	Y	Y	Y
- Implementation of IV is practical and cost effective	Y	Y	Y	Y	Y
- IV implemented over extended time periods, by typical agents, in typical context	Y	Y	Y	Y	Y
Indicators met:	19/21	20/21	20/21	21/21	21/21
Categories met:	5/7	6/7	6/7	6/7	7/7

Quality Indicators	Courtade et al. 2010	Courtade et al. 2010 (science only)	Fetko et al. 2013	Fetko et al. 2013 (science only)	Heinrich et al. 2016
1. Participants and Setting					
- Participants described	Y	Y	Y	Y	Y
- Selection described	Y	Y	Y	Y	Y
- Setting described	Y	Y	Y	Y	Y
2. Dependent Variable (DV)					
- DV described	Y	Y	N	N	N
- Quantifiable index	Y	Y	Y	Y	Y
- DV measurement described	Y	Y	Y	Y	Y
- DV measured repeatedly	Y	Y	Y	N	Y
- Inter-observer agreement data reported	Y	Y	Y	N	Y
3. Independent variable (IV)					
- IV described	Y	Y	Y	Y	Y
- IV systematically manipulated	Y	Y	Y	Y	Y
- Procedural fidelity data reported	Y	Y	Y	Y	Y
4. Baseline					
- DV repeatedly measured prior to IV implementation	Y	Y	Y	Y	Y
- Baseline procedures described	Y	Y	Y	Y	Y
5. Experimental Control					
- Three demonstrations of experimental control	Y	Y	Y	N	Y
- Design controlled for common threats to internal validity	Y	Y	Y	Y	Y
- Pattern of results demonstrates experimental control	Y	Y	Y	Y	Y
6. External validity					
- Experimental effects replicated across participants, setting, or materials	Y	Y	Y	Y	Y
7. Social validity					
- DV is socially important	Y	Y	Y	Y	Y
- Magnitude of change in the DV from the intervention is socially important	Y	Y	N	N	Y
- Implementation of IV is practical and cost effective	Y	Y	Y	Y	Y
- IV implemented over extended time periods, by typical agents, in typical context	Y	Y	Y	Y	Y
Indicators met:	21/21	21/21	19/21	16/21	20/21
Categories met:	7/7	7/7	5/7	4/7	6/7

Quality Indicators	Heinrich et al. 2016 (science only)	Hudson et al. 2014	Jameson et al. 2007	Jameson et al. 2007 (science only)	Jimenez et al. 2009
1. Participants and Setting					
- Participants described	Y	Y	Y	Y	Y
- Selection described	Y	Y	Y	Y	Y
- Setting described	Y	Y	Y	Y	Y
2. Dependent Variable (DV)					
- DV described	N	Y	Y	Y	Y
- Quantifiable index	Y	Y	Y	Y	Y
- DV measurement described	Y	Y	Y	Y	Y
- DV measured repeatedly	Y	Y	Y	Y	Y
- Inter-observer agreement data reported	Y	Y	Y	Y	Y
3. Independent variable (IV)					
- IV described	Y	Y	Y	Y	Y
- IV systematically manipulated	Y	Y	Y	Y	Y
- Procedural fidelity data reported	Y	Y	Y	Y	Y
4. Baseline					
- DV repeatedly measured prior to IV implementation	Y	Y	Y	Y	Y
- Baseline procedures described	Y	Y	Y	Y	Y
5. Experimental Control					
- Three demonstrations of experimental control	N	Y	Y	Y	Y
- Design controlled for common threats to internal validity	Y	Y	Y	Y	Y
- Pattern of results demonstrates experimental control	Y	Y	Y	Y	Y
6. External validity					
- Experimental effects replicated across participants, setting, or materials	N	Y	Y	N	Y
7. Social validity					
- DV is socially important	Y	Y	Y	Y	Y
- Magnitude of change in the DV from the intervention is socially important	Y	Y	Y	Y	Y
- Implementation of IV is practical and cost effective	Y	Y	Y	Y	Y
- IV implemented over extended time periods, by typical agents, in typical context	Y	Y	Y	Y	Y
Indicators met:	18/21	21/21	21/21	20/21	21/21
Categories met:	4/7	7/7	7/7	6/7	7/7

Quality Indicators	Jimenez et al. 2012	Jimenez et al. 2014	Johnson et al. 2004	Johnson et al. 2004 (science only)	Karl et al. 2013
1. Participants and Setting					
- Participants described	Y	Y	N	N	Y
- Selection described	Y	Y	N	N	Y
- Setting described	Y	Y	Y	Y	Y
2. Dependent Variable (DV)					
- DV described	Y	Y	Y	Y	N
- Quantifiable index	Y	Y	Y	Y	Y
- DV measurement described	Y	Y	Y	Y	Y
- DV measured repeatedly	Y	Y	Y	Y	Y
- Inter-observer agreement data reported	Y	Y	Y	Y	Y
3. Independent variable (IV)					
- IV described	Y	Y	Y	Y	Y
- IV systematically manipulated	Y	Y	Y	Y	Y
- Procedural fidelity data reported	Y	Y	Y	Y	Y
4. Baseline					
- DV repeatedly measured prior to IV implementation	Y	Y	Y	Y	Y
- Baseline procedures described	Y	Y	Y	Y	Y
5. Experimental Control					
- Three demonstrations of experimental control	Y	Y	Y	Y	Y
- Design controlled for common threats to internal validity	Y	Y	Y	Y	Y
- Pattern of results demonstrates experimental control	Y	Y	Y	Y	Y
6. External validity					
- Experimental effects replicated across participants, setting, or materials	Y	Y	Y	Y	Y
7. Social validity					
- DV is socially important	Y	Y	Y	Y	Y
- Magnitude of change in the DV from the intervention is socially important	Y	N	Y	Y	Y
- Implementation of IV is practical and cost effective	Y	Y	Y	Y	Y
- IV implemented over extended time periods, by typical agents, in typical context	Y	Y	Y	Y	Y
Indicators met:	21/21	20/21	19/21	19/21	20/21
Categories met:	7/7	6/7	6/7	6/7	5/7

Quality Indicators	Karl et al. 2013 (science only)	Knight et al. 2012	Knight et al. 2013	Knight et al. 2014	Knight et al. 2017
1. Participants and Setting					
- Participants described	Y	Y	Y	Y	Y
- Selection described	Y	Y	Y	Y	Y
- Setting described	Y	Y	Y	Y	Y
2. Dependent Variable (DV)					
- DV described	N	Y	Y	Y	Y
- Quantifiable index	Y	Y	Y	Y	Y
- DV measurement described	Y	Y	Y	Y	Y
- DV measured repeatedly	Y	Y	Y	Y	Y
- Inter-observer agreement data reported	Y	Y	Y	Y	Y
3. Independent variable (IV)					
- IV described	Y	Y	Y	Y	Y
- IV systematically manipulated	Y	Y	Y	Y	Y
- Procedural fidelity data reported	Y	Y	Y	Y	Y
4. Baseline					
- DV repeatedly measured prior to IV implementation	Y	Y	Y	Y	Y
- Baseline procedures described	Y	Y	Y	Y	Y
5. Experimental Control					
- Three demonstrations of experimental control	Y	Y	Y	Y	Y
- Design controlled for common threats to internal validity	Y	Y	Y	Y	Y
- Pattern of results demonstrates experimental control	Y	Y	Y	Y	Y
6. External validity					
- Experimental effects replicated across participants, setting, or materials	Y	Y	Y	Y	Y
7. Social validity					
- DV is socially important	Y	Y	Y	Y	Y
- Magnitude of change in the DV from the intervention is socially important	Y	Y	Y	N	Y
- Implementation of IV is practical and cost effective	Y	Y	Y	Y	Y
- IV implemented over extended time periods, by typical agents, in typical context	Y	Y	Y	Y	Y
Indicators met:	20/21	21/21	21/21	20/21	21/21
Categories met:	6/7	7/7	7/7	6/7	7/7

Quality Indicators	McDonnell et al. 2006	McDonnell et al. 2006 (science only)	McMahon et al. 2016	Miller and Taber-Doughty 2014	Miller et al. 2015
1. Participants and Setting					
- Participants described	N	N	Y	Y	Y
- Selection described	Y	Y	Y	Y	Y
- Setting described	Y	Y	Y	Y	Y
2. Dependent Variable (DV)					
- DV described	Y	Y	Y	Y	Y
- Quantifiable index	Y	Y	Y	Y	Y
- DV measurement described	Y	Y	Y	Y	Y
- DV measured repeatedly	Y	Y	Y	Y	Y
- Inter-observer agreement data reported	Y	Y	Y	Y	Y
3. Independent variable (IV)					
- IV described	Y	Y	Y	Y	Y
- IV systematically manipulated	Y	Y	Y	Y	Y
- Procedural fidelity data reported	Y	Y	Y	Y	Y
4. Baseline					
- DV repeatedly measured prior to IV implementation	Y	Y	Y	Y	Y
- Baseline procedures described	Y	Y	Y	Y	Y
5. Experimental Control					
- Three demonstrations of experimental control	Y	Y	Y	Y	Y
- Design controlled for common threats to internal validity	Y	Y	Y	Y	Y
- Pattern of results demonstrates experimental control	Y	Y	Y	Y	Y
6. External validity					
- Experimental effects replicated across participants, setting, or materials	Y	Y	Y	Y	Y
7. Social validity					
- DV is socially important	Y	Y	Y	Y	Y
- Magnitude of change in the DV from the intervention is socially important	Y	Y	Y	Y	Y
- Implementation of IV is practical and cost effective	Y	Y	Y	Y	Y
- IV implemented over extended time periods, by typical agents, in typical context	Y	Y	Y	Y	Y
Indicators met:	20/21	20/21	21/21	21/21	21/21
Categories met:	6/7	6/7	7/7	7/7	7/7

Quality Indicators	Riesen et al. 2003	Riesen et al. 2003 (science only)	Riggs et al. 2013	Smith et al. 2013a	Smith et al. 2013b
1. Participants and Setting					
- Participants described	N	N	N	Y	N
- Selection described	Y	Y	Y	Y	Y
- Setting described	Y	Y	Y	Y	Y
2. Dependent Variable (DV)					
- DV described	Y	Y	N	Y	Y
- Quantifiable index	Y	Y	Y	Y	Y
- DV measurement described	Y	Y	Y	Y	Y
- DV measured repeatedly	Y	Y	Y	Y	Y
- Inter-observer agreement data reported	Y	Y	Y	Y	Y
3. Independent variable (IV)					
- IV described	Y	Y	Y	Y	Y
- IV systematically manipulated	Y	Y	Y	Y	Y
- Procedural fidelity data reported	Y	Y	Y	Y	Y
4. Baseline					
- DV repeatedly measured prior to IV implementation	Y	Y	Y	Y	Y
- Baseline procedures described	Y	Y	Y	Y	Y
5. Experimental Control					
- Three demonstrations of experimental control	Y	Y	Y	Y	Y
- Design controlled for common threats to internal validity	Y	Y	Y	Y	Y
- Pattern of results demonstrates experimental control	Y	Y	Y	Y	Y
6. External validity					
- Experimental effects replicated across participants, setting, or materials	Y	Y	Y	Y	Y
7. Social validity					
- DV is socially important	Y	Y	Y	Y	Y
- Magnitude of change in the DV from the intervention is socially important	Y	Y	Y	Y	Y
- Implementation of IV is practical and cost effective	Y	Y	Y	Y	Y
- IV implemented over extended time periods, by typical agents, in typical context	Y	Y	Y	Y	Y
Indicators met:	20/21	20/21	19/21	21/21	20/21
Categories met:	6/7	6/7	5/7	7/7	6/7

Table S3

Quality appraisal scores of studies incorporating group design.

CASP questions	**References** Browder at el. 2010	Roberts and Joiner 2007
(A) Are the results of the trial valid?		
1. Did the trial address clearly focused issue?	Yes	Yes
2. Was the assignment of patients to treatment randomised?	Yes	N/A – non-randomised controlled trial design
3. Were all of the patients who entered the trial properly accounted for at its conclusion?	Yes	Yes
4. Were patients, health workers and study personnel 'blind' to treatment?	No	No
5. Were the groups similar at the start of the trial?	Yes	Yes
6. Aside from the experimental intervention, were the groups treated equally?	Yes	Yes
(B) What are the results?		
7. How large was the treatment effect?	Mathematic group had 27.9% gain at math post-test compared to pre-test and 2.9% gain at science post-test compared to pre-test. Science group had 1% gain at math post-test compared to pre-test and 15.7% gain at science post-test compared to pre-test.	Difference in pre- and post-test results were greater for the concept mapping condition than the conventional teaching condition. Students in conventional teaching condition had 9.4 score increase at post-test compared to pre-test at questionnaires and 14.1 increase in concept map scores. Students in concept mapping condition had 35.6 score increase at post-test compared to pre-test at questionnaires and 33 increase in concept map scores.
8. How precise was the estimate of the treatment effect?	$P<.001$. Math group had much higher gains at math post-test than science group (Cohen's $d = 2.41$). Science group had much higher gains at science post-test than math group (Cohen's $d = 1.33$).	$P<0.05$. The effect size was large ($r=0.66$).
(C) Will the results help locally?		

9. Can the results be applied in your context? (or to the local population?)	Yes	Yes
10. Were all clinically important outcomes considered?	Yes	Yes
11. Are the benefits worth the harms and costs?	Yes	Yes

Table S4

Summary table of systematic instruction interventions.

Source and Origin	Intervention
Browder et al. 2010 USA	Task analyzed inquiry-based instruction. The intervention consisted of three main components: inquiry-based lessons, training targeting science vocabulary, and experiments. Vocabulary was taught using a time delay procedure (involving delivery of the prompt after a specific amount of time after the instruction, usually starting at zero seconds and systematically increasing the interval). The teacher used a range of materials related to the topic of the lesson and engaged students in hands-on experiments while introducing key concepts. All lessons were task analyzed (breaking down a complex task into smaller steps) and conducted by special education teachers in a self-contained classroom.
Collins et al. 2007 USA	Compared three interventions: 1. Simultaneous prompting (the prompt is delivered straight after the instruction and then gradually faded out; controlling probes are conducted before the training to determine if the skills have been acquired) with massed trial instruction (trials are conducted one after the other, without a break in-between). The intervention was delivered by a special education teacher in a resource room. 2. Simultaneous prompting with distributed trial instruction (trials are naturally distributed in daily activities to encourage generalization). The intervention was delivered by special education teacher, instructional assistant, or a peer tutor in general education setting. 3. Embedded instruction (embedded instruction means that the trials are naturally distributed across the sessions and occur as part of students' ongoing routines). The intervention was delivered by an instructional assistant or a peer tutor in a general education classroom.
Collins et al. 2011 USA	Constant time delay procedure (procedure involving delivery of the prompt after a specific amount of time after the instruction, usually starting at zero seconds and then increasing the interval to a specific number of seconds for the rest of the trials) was used to teach the properties of elements. The instructor used a 0-second delay during the first session and a 3-second delay during the consecutive sessions. The prompts were either a verbal model or a verbal model with a gesture.
Collins et al. 2017 USA	Simultaneous prompting procedure (see above for definition) was used to embed core content related to photosynthesis in teaching a practical skill (plant care). The plant care activity was task-analyzed and core content was delivered as part of instructive feedback after completing plant care steps. No response was required of the students. After the intervention phase finished, students were taught photosynthesis content that they had not acquired previously using a simultaneous prompting procedure.
Courtade et al. 2010 USA	Multi-component training in task analyzed inquiry-based instruction for the teachers. The training included: a fidelity checklist, training manual, verbal explanation of content, video modelling, and time to develop one lesson and receive feedback from the researchers. The training was delivered in a one-to-one setting by a researcher and lasted 4 hours. The teachers were also trained in using the system of least-to-most prompts (hierarchy of prompts used to help the students, starting from the least intrusive) error correction and reinforcement.
Fetko et al. 2013 USA	Simultaneous prompting procedure with core content (science vocabulary) embedded as non-target information while teaching a leisure skill activity (UNO game). Peer tutors taught a task-analyzed UNO game to students with disabilities using the simultaneous prompting procedure. The core content (science vocabulary) was delivered after praise for completing each step of the task analysis as part of the instructive feedback.

Heinrich et al. 2016 USA	Embedded simultaneous prompting procedure (see above for definitions). The intervention was delivered by paraprofessionals and peer tutors and took place during several points of the day. Controlling probes to check students' progress were conducted daily before the start of the session.
Hudson et al. 2014 USA	Peer-delivered system of least prompts with adopted science read-alouds – Prior to the start of the intervention peer tutors were trained in the teaching procedure (system of least prompts; see above for the definition) and participants were trained to request help and in the use of self-monitoring tools. The intervention was delivered in a one-to-one format. During each session, the peer tutor read science related text while stopping at predisposed points and asking one of six comprehension questions. If the participant requested help the peer tutor delivered the next step of the predetermined prompting hierarchy. If the participant did not respond or responded incorrectly the peer tutor delivered the correction procedure.
Jameson et al. 2007 USA	Comparison of two interventions: 1. One-to-one embedded instruction (see above for the definition) – implemented by the special education teacher and a paraprofessional in the general education class. The intervention trials were delivered during transitions, breaks, etc. The procedure also involved constant time delay (see above for definition), differential reinforcement (procedure involving rewarding independent correct response and withholding reward when prompt is needed), and error correction. 2. One-to-one massed trials instructional format (see above for definition) - implemented by the special education teacher and a paraprofessional in the self-contained special education class. The same procedures were used in the mass trial condition as in embedded instruction. The main difference was that the trials were staggered together and delivered during one session per day one after the other without any pause in between.
Jimenez et al. 2009 USA	Multicomponent training package for students – the package consisted of multiple exemplar training (procedure involving teaching a target instruction across different materials, settings or people at the same time to facilitate generalization), time delay (see above for definition) and self-directed learning prompts (KWHL chart - What we know?; What we want to know?; How to find out?; What was learned?). The training occurred in a one-to-one setting and was delivered by a researcher. Students were taught to turn pages of the workbook, state their response, and complete the KWHL chart to facilitate self-directed learning. Students' generalization of the use of the KWHL chart was assessed during general education classes.
Jimenez et al. 2012 USA	Peer-mediated embedded instruction with time delay (see above for definitions) – during each lesson peer tutors trained participants on science responses using time delay and embedded instruction and on the use of a KWHL chart using embedded instruction. The intervention took place in the general education classroom and was delivered by peers without disabilities who received one-hour training prior to the start of the study. The science teacher delivered instruction for the whole class first and then peer tutors delivered the teaching trials one-to-one to the participants.
Jimenez et al. 2014 USA	Inquiry-based curriculum for students with severe disabilities – Early Science Curriculum – was implemented across two experimental conditions: 1. Scripted lessons (a detailed script outlining what the teacher needs to say, the teaching procedures to be used, and the order in which the lesson has to progress) – the teacher delivered the content covered in the Early Science Curriculum script using a range of systematic instruction procedures such as time delay (see above for definition), system of least-to-most prompts (see above for definition), specific praise (clearly labelling behavior that the child is being praised for) and an example/nonexample procedure (procedure involving presenting the child with an example and nonexample of a target item while clearly labelling: 'This is….' or 'This is not…'). A KWHL chart was also used. All three students were taught in one group.

	2. <u>Scripted lessons with guided notes</u> – the teaching procedure was the same as outlined above with an exception of the inclusion of guided notes for the participants to help retention of key concepts. These materials included printed notes with symbols and appropriate space for the students to insert picture or vocabulary cards.
Johnson et al. 2004 USA	<u>Embedded instruction</u> (see above for definition) – instructional procedures used were: constant time delay (see above for definition), error correction and reinforcement. Initially a zero second delay was used. Later the delay was increased to four seconds. The intervention was delivered by the teacher in the general education classroom.
Karl et al. 2013 USA	<u>Simultaneous prompting procedure</u> (see above for definition) was used to teach science core content within a functional activity. Students had daily cooking sessions with embedded core content training (science, math and reading). The intervention was delivered in a small group format by a teacher.
Knight et al. 2012 USA	<u>Explicit instruction</u> (errorless teaching procedure focused on teaching the student to recognize examples and non-examples) - the teaching procedure involved a model-lead-test strategy (three step teaching procedure involving the teacher modelling the response for the student first, then doing it with the student and then testing student's understanding) with the teacher waiting for student's response for 3 seconds during the final test. The intervention was delivered in a one-to-one setting by the researcher.
Knight et al. 2013 USA	<u>Systematic instruction</u> <u>treatment package</u> which consisted of a constant time delay procedure (see above for definition), an example/non-example procedure (see above for definition) and graphic organizers (visual display that helps with organizing key concepts and facts). Initially a 0-second delay was used and was later increased to a 5-second delay. The intervention was delivered in a one-to-one setting by the researcher.
Knight et al. 2014 USA	<u>Book Builder</u> (BB; software that allows teachers to create their own eTexts/digital books) implemented across three phases: 1. <u>BB only</u> – the software was used on its own with embedded resources, such as hyperlinks, and coaches delivering prompts. 2. <u>BB and explicit instruction</u> (see above for definition) – the procedure was the same as in phase 1 but the coaches delivered explicit prompting (model-lead-test) and students were provided with examples and non-examples of key vocabulary and concepts. 3. <u>BB, explicit instruction and referring to definition</u> – the procedure was the same as in phase 2 with one exception – the coaches provided students with reasoning about why one item was an example and the other a non-example by referring student back to the definition.
Knight et al. 2017 USA	<u>Modified Book Builder</u> (see above for the definition) – the procedure included embedded animated coaches delivering: model-lead-test, examples and non-examples of key concepts and vocabulary (including coaches providing reasons why one item was an example and the other a non-example), and referrals to the definitions. Additionally, students were required to verbally refer to the definitions. The intervention was implemented by the teacher in the classroom setting.
McDonnell et al. 2006 USA	Comparison of two interventions: 1. <u>One-to-one embedded instruction</u> (see above for definition) – The teaching procedures involved constant time-delay, differential reinforcement, and error correction. The trials were implemented during transitions and breaks. The intervention was implemented by a paraprofessional in a one-to-one format in a general education classroom. 2. <u>Small-group spaced-trial instruction</u> (procedure involved delivering teaching trials to individual students with short breaks or with an activity in-between) – The teaching procedures involved constant time-delay, differential reinforcement, and error correction. The trials were presented to

	students individually in turns. The intervention was implemented by a paraprofessional in a small group format (target pupils and two peers) in a self-contained special education classroom.
Riesen et al. 2003 USA	Embedded instruction (see above for definition) with comparison between: 1. Constant time delay (see above for definition) – Initially a 0-second delay was implemented. After the student correctly defined all target words two for two consecutive times, a 3-second delay was introduced. Error correction was implemented for incorrect responses. The intervention was delivered by paraprofessionals in the general education class during transitions and breaks. 2. Simultaneous prompting (see above for definition) – One test trial was always presented before prompted trials. The correct response was always modelled straight after the instruction. Error correction was implemented for incorrect responses. The intervention was delivered by paraprofessionals in the general education class during transitions and breaks.
Riggs et al. 2013 USA	Constant time delay procedure with examples and non-examples (see above for definitions) – A 0-second time delay was used during the first session and 5-second delay during following sessions. Error correction was implemented for incorrect responses. The intervention was delivered by a special education teacher in a one-to-one or small group format in a special education classroom.
Smith et al. 2013a USA	Embedded computer-assisted explicit instruction (see above for definitions) – An iPad was used to deliver the intervention in a model-test explicit instruction format. The intervention was implemented by a researcher in a one-to-one format in a general education classroom during students' independent study time.
Smith et al. 2013b USA	Task analyzed science inquiry lessons – Early Science Curriculum – The curriculum included scripted lessons, task analyses, explicit instruction (see above for definition), and practical activities/experiments. The intervention was delivered by a teacher in a group format in the students' usual classroom.

Figure 1. A flow diagram illustrating study selection process (adapted from PRISMA Diagram – Moher et al., 2009).