Title	How reliable is assessment of children's sentence comprehension using a self-directed app? A comparison of supported versus independent use
Authors	Frizelle, Pauline;Buckley, Ana;Biancone, Tricia;Ceroni, Anna;Dahly, Darren L.;Fletcher, Paul;Bishop, Dorothy V. M.;McKean, Cristina
Publication date	2023-08-14
Original Citation	Frizelle, P., Buckley, A., Biancone, T., Ceroni, A., Dahly, D., Fletcher, P., Bishop, D.V.M. and Mckean, C. (2023) 'How reliable is assessment of children's sentence comprehension using a self-directed app? A comparison of supported versus independent use', Journal of Child Language, (29 pp). https://doi.org/10.1017/S0305000923000545
Type of publication	Article (peer-reviewed)
Link to publisher's version	https://doi.org/10.1017/S0305000923000545 - 10.1017/s0305000923000545
Rights	© The Author(s), 2023. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited. - http://creativecommons.org/licenses/by/4.0
Download date	2025-04-19 04:25:06
Item downloaded from	https://hdl.handle.net/10468/15014

University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Journal of Child Language (2023), 1–29
doi:10.1017/S0305000923000545

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

How reliable is assessment of children's sentence comprehension using a self-directed app? A comparison of supported versus independent use

Pauline FRIZELLE[1], Ana BUCKLEY[1], Tricia BIANCONE[1], Anna CERONI[1], Darren DAHLY[1], Paul FLETCHER[1], Dorothy V. M. BISHOP[2] and Cristina MCKEAN[3]

[1]Department of Speech and Hearing Sciences, University College Cork, Republic of Ireland
[2]Department of Experimental Psychology, University of Oxford, UK
[3]Newcastle University, UK
Corresponding author: Pauline Frizelle; Email: pp.frizelle@ucc.ie

(Received 30 November 2022; revised 21 August 2023; accepted 30 August 2023)

Abstract

This study reports on the feasibility of using the Test of Complex Syntax- Electronic (TECS-E), as a self-directed app, to measure sentence comprehension in children aged 4 to 5 ½ years old; how testing apps might be adapted for effective independent use; and agreement levels between face-to-face supported computerized and independent computerized testing with this cohort. A pilot phase was completed with 4 to 4;06-year-old children, to determine the appropriate functional app features required to facilitate independent test completion. Following the integration of identified features, children completed the app independently or with adult support (4–4;05 (n = 22) 4;06–4;11 months (n = 55) and 5 to 5;05 (n = 113)) and test re-test reliability was examined. Independent test completion posed problems for children under 5 years but for those over 5, TECS-E is a reliable method to assess children's understanding of complex sentences, when used independently.

Keywords: complex syntax; comprehension; online assessment; children

Introduction

Speech and language therapists carry out language assessments to achieve a range of objectives. These include initial screening and differential diagnosis; establishing standard scores to determine eligibility for services; identifying targets for intervention; and measuring progress in treatment. For the purposes of research, language assessments are used to establish baseline and outcome measurements, to detail phenotyping for genetic studies and for characterising populations and drivers of individual differences in epidemiological studies. (Tomblin et al., 1996; Paul & Norbury, 2012). Although therapists are trained to use multiple assessment techniques and approaches, such as language

sampling; validated parent reports; observational methods; and holistic profiling, standardized assessments are extensively used (Betz et al., 2013; Hoffman et al., 2011). Betz et al. (2013) found that 50% of SLTs working with school aged children rated standardized tests to be the most important diagnostic assessment measure. Additionally, they are the most used measure, by researchers, to identify study participants with developmental language disorder (DLD) (Betz et al., 2013). Standardized language assessments are designed to be administered reliably and consistently on repeated occasions. In developmental work, they provide standard scores, and percentile ranks in relation to children of a similar age from a normative sample.

Most standardized assessments in the field of speech and language therapy are currently delivered in person and involve children interacting with toys or pictorial materials. However, in recent years there has been a move towards developing computerized assessments. Computerized testing is far from standard practice and these tests are often administered in person, with varying levels of examiner support, particularly when assessing young children. In such cases, the primary advantage of the computerized format is that it allows for the presentation of dynamic materials, such as videos, and can automate scoring to reduce examiner error. However, there could be benefits from having computerized tests that can be self-directed i.e., completed independently by children. This would facilitate test administration in groups and potential online administration. Internet-ready assessments could integrate the benefits of computerized testing with the advantages afforded by the world wide web, namely a fast and cost-effective method of assessing large study populations. To the best of our knowledge no study has reported on the feasibility of using independent language comprehension testing with preschool/young school aged children; how these tests might be adapted for effective, independent, and therefore potential group or online[1] use; or agreement levels between supported computerized and independent computerized testing with this population. This is the focus of the current study.

Computerized assessments

Assessing comprehension is particularly challenging and is usually measured by observing a child's behavior in response to language. Computerized assessments allow for the assessment of language constructions that would be very difficult to assess without the use of technology (such as complex sentences including mental state verbs – see Frizelle et al., 2019a). Traditional assessment methods often rely on multiple-choice picture selection tasks which place a significant cognitive and memory burden on the child (Frizelle et al., 2017b, 2019a). Using animations, computerized assessments allow testing of children's comprehension in real time. While still images require children to infer movement in characters and temporal relationships (which are often depicted very subtly), the use of animations can depict actions two or three dimensionally. Bringing characters to life in real time reduces the level of inference making required by the child, while at the same time potentially increasing levels of engagement.

The use of computerized assessments may also facilitate the development of standardized tests with better psychometric properties – the quality of which are a cause for concern in many tests that are currently commercially available (Spaulding et al., 2006;

[1]Note – our use of the term online throughout this paper refers to independent test completion online rather than virtual SLT administration.

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

Denman et al., 2017). A robust test should be valid and reliable with a high degree of diagnostic accuracy. Computerized tests have been shown to allow for improved standardization in cognitive and neuropsychological testing (Barak & English, 2002; Bilder, 2011; Parsey & Schmitter-Edgecombe, 2013) which is one aspect of test reliability. In paper and pencil based assessments testers are reminded in test manuals to be aware of their speech rate, loudness and general style of test delivery; however variation between testers is inevitable. In addition, both language impaired and typically developing children have been shown to have greater understanding of sentences (with a medium effect size), when spoken in an accent that they were familiar with, when compared to an unfamiliar one (Frizelle et al., 2017a) – a point that is rarely highlighted as any cause for concern in the assessment process. In computerized testing both the test instructions and items being assessed can be pre-recorded in a supra-regional accent associated with the country in which the test is being taken. Computerized testing therefore has the advantage of more precise control of test instructions and stimulus items, with consistency of accent or prosodic features such as rate or stress. Given that computerized testing can involve automated presentation and recording of children's responses, this opens the way for those with less training in assessment administration to provide test supervision, without negatively impacting on reliability (Carson et al., 2011). Automated scoring is also advantageous in that it is less sensitive to errors compared to when results are manually entered (see Kraut et al., 2004; Naglieri et al., 2004 with respect to psychological testing). Measures such as response time can also be integrated into computerised testing (Barak & English, 2002; Bauer et al., 2012; Bilder, 2011; Naglieri et al., 2004) and increase sensitivity to more subtle impairments and individual differences. Studies have shown that where there are no differences in children's test performance in terms of language comprehension accuracy, differences may be present in response time (Kalff et al., 2005). In addition, through multi-media elements and dynamic animations, feedback can be given in controlled and innovative ways. Previous literature suggests that type and rate of feedback can influence children's test performance within a standardized assessment context (Shute, 2008); through computerized testing we can control this variable.

Online / independent testing

The process of acquiring an appropriate normative sample is central to standardized test validity. However, it is expensive to recruit samples that are unbiased and of adequate size and diversity. Standard language scores and percentile ranks are invalid if the sample on which they are based is not sufficiently large to allow for valid comparisons; is not balanced with respect to gender and age or does not include children from different socio-economic backgrounds. The use of online testing or simultaneous testing in large groups have significant advantages in this regard as they allow for increased accessibility to large population-based samples, who can be assessed in a more cost effective and efficient manner (Bauer et al., 2012; Naglieri et al., 2004). Other advantages include automatic scoring which allows test scores to be added to a test's database so that norms can be adjusted accordingly. As quick inexpensive methods for collecting data on large and potentially diverse samples, online or group testing are particularly valuable resources in research. Haworth et al. (2007) estimate that individual in-person cognitive testing (in the home) costs an average of about £170 per test session in the UK and highlight that this is unsustainable when large samples are required or for longitudinal research. In addition, research on questionnaire studies by Gosling et al. (2004) has shown

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

that online samples are generally more diverse with respect to socio-economic status, age and gender, than samples recruited using more traditional methods. However, methods of online sampling can vary and a recent study by Chan et al. (2022) found that recruiting within schools yielded a more racially diverse sample than recruiting parents online. Within research, efficiency is paramount and time frames for testing can often be short, for example in large scale randomized control intervention trials. When constrained by time, traditional one-to-one methods limit both the scale and speed of data collection. The ability to take a test at home or from school may also increase the likelihood that parents with busy lifestyles will consent to their children taking part in research studies (Birnbaum, 2004; Germine et al., 2012; Naglieri et al., 2004). Finally, through reducing the need to travel to specific settings, online testing can facilitate equity of intervention access, in that those living in remote locations or with mobility problems can undergo at least an initial level of assessment before engaging with tele-health interventions.

Challenges

Despite the multiple advantages of online based/computerized testing there are a number of disadvantages. Firstly, the format of online testing results in a lack of control over the testing environment and exposes a child to potential influence by family members in relation to their responses. While recommendations can be made with respect to levels of parental or family involvement and regarding the ideal environment in which the test should be taken, there is a risk that the recommendations will not be adhered to (Yue et al., 2021). On the other hand, online testing may reduce anxiety or embarrassment that may be present when individuals are tested in person (Birnbaum, 2004; Kraut et al., 2004). In any case, we cannot assume that the psychometric properties of an online test are identical to its traditional equivalent (Buchanan, 2003; Buchanan et al., 2010). Even when a typically examiner-administered test is programmed for computer administration it becomes a new and different test (Bauer et al., 2012). Therefore, tests used independently or online require an independent evaluation of their psychometric properties. Using online testing we also have less control over the device a child uses to take the test, for example the screen size; levels of resolution etc., the impact of which may result in test items appearing differently (Bartram, 2006).

In examiner-supported approaches, the child interacts with a person who presents live or pre-recorded stimuli, records the child's verbal or non-verbal responses and makes note of key behavioural / communicative observations. Independent online testing does not include any element of human interaction. Mode of response can also vary between examiner supported and online testing (even when both methods involve the use of a computer). Whereas pointing or speaking may suffice in person, online testing will differ between tablet and desktop computers (touch screen vs mouse click). Speed of internet connection may also cause variation in the duration of a test and loss of connection may terminate the assessment altogether (with an associated loss of data). In summary, technology-based assessments have the potential to solve a number of problems with respect to standardized tests in language, cognition and general educational attainment; however, their use online is not without challenges. Despite these challenges, the consensus in the literature, is that online testing is a feasible method of assessment for older children (Feenstra et al., 2018; Haworth et al., 2007). Research regarding technology-based assessment in childhood has focussed on differences between in-person and computerized test performance with respect to validity and reliability, advantages and

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

disadvantages and the effects of other variables, such as gender (Gallagher et al., 2000; Hamann et al., 2016). In general, findings suggest that older children's scores (late primary early secondary school age) are not significantly affected by the modality of delivery. For example, Haworth et al. (2007) compared online and standard paper and pencil versions of reading and maths tests in thirty 12-year-old children and found correlations of .80 between the two modes of delivery.

Until recently, little attention had been given to technology-based testing of children under the age of 12 (Carson et al., 2011). In studies where a paper-based test is compared to its computerised equivalent, language has rarely been the domain of focus. Early work by Maguire et al. (1991) compared elementary school children's expressive vocabulary results on a paper based versus computerised version of the Peabody Picture vocabulary test. Although no differences were reported, both test versions involved a level of examiner support and the only distinction between both versions was the mode of response (keyboard versus pointing response). Carson et al. (2011) investigated whether a computer-based administration of a phonological awareness test would yield similar results to a paper-based administration facilitated by an examiner, with 4- and 5-year-old children. Two-thirds of the group (n= 21) had typical speech and language development and one third (n=12) had a moderate speech delay. Interestingly both tests generated comparable scores for all children, with the computer-based method taking 20% less time to administer.

In contrast, Csapó et al. (2014) compared face to face administered and computerized versions of four tests of school readiness, with a large group of children at school entry. The tests included speech sound discrimination, relational reasoning (the ability to understand the meaning of words that are related), counting and basic numeracy, and deductive reasoning. Differences in how children performed between delivery modes depended on the specific test, with the online versions of speech sound discrimination, relational reasoning and deductive reasoning showing increased reliability relative to face to face delivery. Overall, they found that children's performance was lower on computer-based tests than on the face-to-face equivalents. Csapó et al. suggest that this may be due to teachers giving children higher scores than when scores are automatically, and therefore more objectively, assigned. But they suggest that it could also reflect children having difficulty navigating computerized tests. One of the difficulties interpreting the findings in relation to computerized test versions is that the presence or absence of a test supervisor is often not explicitly stated and this is problematic as the level of independence in test taking is therefore unclear.

More recently, Lo et al. (2021) compared an in-laboratory and online (at home) administration of a computerised touchscreen recognition task (two alternative forced choice) with 18- 20 month old toddlers and found no differences in children's performance on the task in each setting. However, both administrations were supported in that parents/ experimenters were required to tap on the 'Next' button at different stages throughout the assessment.

Administering computerized tests online to children at preschool and young school age raises a number of questions concerning the validity of the results. Depending on how the assessment has been devised, the presence or absence of a level of supervision is likely to be influential – even when comparing two computerized versions of a test (online and in person). To harness the advantages of computerised testing while at the same overcoming many of the challenges, one solution is to develop tablet-based tests for independent use, but to administer them in person, in large groups with supervision. Although children show increasing proficiency using tablets (Marsh et al., 2015), if

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

current assessment tools are to be used independently in this way, certain adaptations would need to be put in place to allow self-directed completion of a given test. Stock et al. (2004) compared existing paper-based assessments with internet-based multi-media testing in a group of adolescents and adults with mild and moderate intellectual disability. Before comparing the two modes of assessment, they initially determined the feasibility of use and necessary functional features of the internet-based assessment, to allow self-directed administration. Interestingly, they found no difference between paper-based and internet-based assessment with respect to the accuracy of the results, but did find a significant difference in the number of prompts required to complete the internet test (2.2) when compared to pen and paper delivery (7.5). Results showed that with the appropriate audio, video and picture supports the internet test allowed for a more independent administration. Additionally they reported that participants enjoyed the control and self-pacing, as well as the multi-media aspects of the computerised test. It is possible that similar supports would allow young children to engage with computerised testing relatively independently. However, we expect that this would very much depend on the age of the child; how the overall test is designed; and the level of computer interaction required to complete the test.

The current study

The current study explores these issues through adapting The Test of Complex Syntax – Electronic (TECS-E) a computerized test of complex sentences, with the aim of developing the tool as an app which could be completed independently. To date TECS-E has been reported on in two previous studies. The first, a methodological comparison study in which it was administered with 104 typically developing children between the ages of 3;06 and 4;11 years (Frizelle et al., 2019a). The second, a group comparison study (using a slightly shorter version of TECS-E) in which children with Down syndrome (n = 33) were compared with those with cognitive impairment of unknown aetiology and typically developing children matched on a non-verbal measure of cognitive ability (Frizelle et al., 2019b). In the latter study Cronbach's alpha (a measure of internal consistency) was calculated for the whole sample, as .877. The first iteration of TECS-E was computerized with the following features: it allowed for an adult to input background information on the participating child; audio test stimuli were pre-recorded; test stimuli were presented through the use of animation; TECS-E was delivered on a computer tablet device (Micro-soft surface Pro); the child interacted with the device by touching the screen to choose their response; there were some animated in-built motivator/ reward items; and scoring was automated. Despite these features, in both previously published studies TECS-E has always been delivered with the support of an examiner. A scripted explanation of how the assessment works, and what was required of the child, was given in person. While progressing through the practice items the examiner's responses were dependant on the child's performance and feedback was not tightly controlled. If the child had difficulty interacting with the device (i.e., moving from one screen to the next or touching an arrow to allow them to hear the test item again) the examiner was at liberty to do this on the child's behalf. With respect to some younger children who pointed to the correct part of the screen to give their response but did not actually touch the screen, the examiner was free to touch the screen for them (while in no way influencing the child's response). The aim was to evaluate the child's understanding of complex sentences as accurately as possible and therefore we

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

did not want to negatively impact children's performance: by any uncertainty around the process or computer interface. In addition, TECS-E had been designed such that examiner support was expected.

There would be significant benefits to having a version of TECS-E that could be completed by children independently, such as screening children at scale; identifying whole class progress; and allowing for test completion in different educational settings without significant levels of staff training in test administration. However, for these to be realised we need to ascertain how TECS-E could be optimally adapted and at what age the results could be judged to be valid and reliable. To the best of our knowledge no study has reported on the feasibility of using a self-directed app to measure sentence comprehension with children from 4 years old; how testing apps might be adapted for effective self-directed use; or agreement levels between face to face supported computerized and self-directed computerized testing with this population. This is the focus of the current study, of which there were two parts.

- The pilot phase was used to consider the feasibility of use and the required functional features of the Test of Complex Syntax- Electronic (TECS-E), to allow self-directed/independent completion for children between the ages of 4 and 4;06 years.
- In the main study, we ask if
  1. Following the implementation of these features, is the agreement and reliability of the independent (self-directed) version of the TECS-E similar to that of the supported test, based on TECS-E overall accuracy scores, in 4 to 4;06-year-old children? Then, if reliability is poor in 4 to 4;06 year olds, can we identify the youngest age group for which the independent delivery of the test will be reliable enough to be used in practice?
  2. Do the two methods of test administration agree in terms of the order of difficulty of specific constructions. If they do, then this would further validate the use of the independent method of assessment as a reliable indication of children's understanding of complex sentences.

Hypotheses

Our pre-registered hypotheses were as follows https://osf.io/emfcy/.

Pilot study

Given our experience using TECS-E with previous typically developing and intellectually disabled populations we anticipated that several new features would be required to facilitate independent test taking. To identify the appropriate features, we planned to monitor children (between 4 and 4;06 years) while carrying out the test in its original form, using a structured observation tool and documenting any instances where examiner support might be required. The process was intended to be iterative, in that, following the implementation of these changes, a new group of children would attempt to complete the test without any adult support. Further features would then be refined / altered accordingly. We assumed that if children aged 4 years could carry out the assessment independently, older children would not have difficulty with the same process. The results did not undergo statistical analysis but are reported narratively.

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

Main study

Hypothesis 1: With the appropriate adaptations put in place, our aim was to allow independent test completion from the age of 4 years. However, given the cognitive ability of children at this young age we hypothesized that the re-test reliability in the accuracy scores between the independent and the examiner supported administration of the TECS-E might be poor. To evaluate this hypothesis, we used a group-sequential testing procedure to, as quickly as possible, identify poor test performance in 4 to 4;06-year-olds, defined as a test-retest *r* lower than 0.75 (described in more detail under *Sample Size Rationale* and S*topping Rules* below). If we found early evidence of poor test performance, we would start recruiting from an older age group (4:06 to 5-year-olds) until the end of the planned recruitment.

At end of recruitment, we planned to evaluate the reliability and agreement of the independent version of the test (vs the supported delivery) and the degree to which participant age predicts test performance (see *Statistical Methods* below). While our final evaluation of the independent performance on the test would rely on the complete analysis, we set our acceptance level of the test-retest r as close to 0.75 or greater. This is in keeping with Csapó et al. (2014) who found a high degree of correlation ($r = 0.75$) between a language based deductive reasoning task administered face-to-face and online, with first grade children in Hungary. It is also consistent with Stock et al. (2004) who, following feasibility and assessment adaptation work, found no difference between modes of assessment, with respect to accuracy in a group of older participants with intellectual disability.

Hypothesis 2: Given that the same items were used in both tests we hypothesized that the same rank ordering of constructions would emerge between the independent and supported test administrations. In a previous study in which two different methods of assessing young children's understanding of relative clauses (sentence repetition and multiple-choice sentence picture matching tasks) were compared, we found that even though there were individual differences in how children performed on both measures, the assessments revealed a similar order of difficulty of constructions (Frizelle et al., 2017b). In this study the same methodology (truth value judgement) is used in both conditions and therefore we anticipate a similar rank ordering of constructions across conditions.

Pilot study

Method

Study design

The pilot study was a qualitative observational study.

Participants

The aim was to recruit 14 children between the ages of 4;0 and 4;06 years (7 on which to base the initial test adaptations and a further 7 to ascertain their effectiveness). Nine children were initially recruited and 2 were subsequently excluded because they observed the test being completed by other children. The included children had a mean age of 50.86 months (SD = 1.46) and 4 were boys. Due to the Covid-19 pandemic it was not possible to recruit 14 children in the initial stages of the work. Therefore, to progress

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

with the study, test alterations were made based on the feedback from the first 7 children.

The age range chosen reflects a stage in development in which most children are not literate but are likely to have accrued some experience using tablet-based computers. Therefore, while they are likely to require significant test adaptations to facilitate independent test-taking, we expect that they are familiar with the medium through which the test is delivered. To account for potential differences in levels of computer exposure, children were recruited through preschools serving communities with varying levels of social dis/advantage, in Cork city in the Republic of Ireland. The parents of participants were required to give written consent and children were asked to sign an assent form. Ethical approval for the study was granted by the Social Research Ethics Committee, University College Cork. Children were included on the basis that they had typical language abilities (based on teacher and parental reports), had never been referred to speech and language therapy, spoke English as their primary language, had no known intellectual or neurological difficulties and no sensory-neural or conductive hearing loss. Children's hearing was screened prior to carrying out the assessment using the Ling six sound test; an assessment deemed to be an appropriate measure of the ability to hear speech at a level commensurate with everyday conversation.

TECS-E complex syntax comprehension task

TECS-E is a measure of complex syntax which has been developed for use on a tablet or computer device. The assessment is an animated sentence comprehension task that uses a truth value judgement paradigm. The advantage of using this type of task where children are shown individual animations is that they can evaluate the truth of each sentence directly against the real world scenario without having to store in memory the arguments associated with the verbs. In this regard the task has no greater memory load than that required for processing language in everyday discourse.

In TECS-E children are shown test animations (approximately 6 seconds in length) each with an accompanying auditory test sentence. The test sentence accurately describes the animation for half of the items and for the remaining items the test sentence and animation are incongruent. At the bottom left and right corners of the screen there is a smiley and sad face. Children are asked if what is shown in the animation matches the sentence they hear and to respond accordingly by touching either the smiley or sad face on the screen. All test sentences are pre-recorded by a native female English speaker. The test animations represent relative clauses, adverbial clauses, and sentential complements. There are an additional 10 catch items designed to identify children who are showing a *yes* bias. Ten motivational star-animations are also integrated into the test. To date TECS-E has been administered where each structure was represented by either 8 or 10 animations (4/5 match and 4/5 non-match respectively) (see Frizelle et al., 2019a, 2019b). In both iterations, animations represented one of 5 types of relative clause (*intransitive subject, transitive subject, object, oblique and indirect object*) four types of adverbial clauses (*before, after, because and if*); and four types of sentential complements (*think, know, pretend and wish*). Based on the British National Corpus all constructions included high frequency nouns and verbs.

Relative clause test sentences were chosen based on previous work carried out by Diessel and Tomasello (2000, 2005) and Frizelle and Fletcher (2014) indicating a performance hierarchy in children's knowledge of these constructions. All were fully

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

bi-clausal and designed to reflect structures that are used in natural discourse. To this end they were attached to the direct object of a transitive clause and object relatives had an inanimate head noun and a pronominal subject (see Kidd et al., 2007).

The design of our task aims to mirror language processing in natural usage. In natural discourse a relative clause is used contrastively, for example, a sentence such as *He found the girl that was hiding* would imply the existence of another girl who is not hiding. We can therefore enhance the ecological validity of an assessment tool by structuring each relative clause item so that there is an alternative to the head noun to which the relative clause refers i.e., a referent from which another can be distinguished. For example the representation of the sentence *The boy picked up the cup that she broke* includes two cups within the scene that fall and are broken. Children are therefore required to make a semantic evaluation of a sentence and to map the thematic roles to the appropriate verb argument structures without having to actively rule out three competing alternative mappings. This is in contrast to the usual multiple choice sentence picture matching task in which the correct representation and three foils/competitors (regarding who did what to whom) would be typically presented. For example, for the sentence *He found the girl that was hiding*, images depicting *She found the boy that was hiding; The boy that was hiding, found the girl;* and *The girl that was hiding, found the boy,* would also be shown. However, our previous work shows that by using this type of approach in which there is a contrastive component for every role, the child is required to actively rule out three competitors in a way that would never happen in the everyday comprehension of language. Our work showed that this multiple choice methodology is artificially elevating the difficulty level of the task and results in an assessment of factors other than those that are linguistic (see Frizelle et al., 2017b). By integrating a distractor within each item and presenting both congruent and incongruent items we aimed to overcome this difficulty (see Frizelle et al., 2019a).

Adverbial clauses were chosen to reflect 1) conjoined conjunctions with a range of frequency use by young children (see Diessel, 2004) and 2) a range of functions (two temporal (*before, after*), one causal (*because*) and one conditional (*if*). Iconicity was controlled for such that in half the adverbial constructions the clause order reflects the order of events in the real world, while in the remainder of items the order is reversed. This will allow us to determine whether a child is using a strategy based on iconicity when trying to interpret each adverbial item.

Finally, sentential complements were chosen so that three of the complement taking verbs were mental state verbs (*think, know, pretend*) and one represented desire (*wish*). These were informed by acquisition data reported by Diessel (2004), again representing a range with respect to how frequently they are used by young children as well as those that could be adequately represented through animation. With respect to the congruent/true items for the *know* constructions one could argue that a child could respond correctly by understanding the simple sentence that follows the *know* verb (e.g., *She knows the boy ate the sweets* (where we see in the animation that he does eat the sweets) and indeed it is the incongruent/false items that give more information about the child's understanding of these constructions. In contrast the reverse is the case for the *think* items, where the congruent items reveal more about the child's understanding. Importantly, it is the pattern of results that the child shows across all items that allows us to profile the child's understanding i.e., they need to get both true and false items correct to show complete understanding.

For the purposes of this study in which the focus was assessment methodology, a shortened version of the tool was used. Sentences were chosen within each family of

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

Table 1. Example Test Sentences

Sentence type	Example sentence	YouTube video link
Relative clause		
Subject intransitive	He found the girl that was hiding.	https://www.youtube.com/watch?v=bA6QCvVs4j4
Object	The boy picked up the cup that she broke.	https://www.youtube.com/watch?v=zsDntFWAhSI
Complement clause		
Think	She thinks the boy's hair is dry.	https://www.youtube.com/watch?v=be-llO4muDg
Know	She knows the man took her dog.	https://www.youtube.com/watch?v=54zowGeQ870
Adverbial clause		
Before	The boy played football before he watched TV.	https://www.youtube.com/watch?v=CMnBOmu5kxE
After	The girl opened the box after she put on her slippers.	https://www.youtube.com/watch?v=HxpviG7iod4

constructions to reflect a semantic contrast and two distinct levels of syntactic difficulty (intransitive subject and object relatives; before and after; think and know) (Frizelle et al., 2019b). Catch item animations depicted simple sentences, considered to be well within the children's developmental level. The design of the non-match item is dependent on the type of structure being assessed. For a detailed description see Frizelle et al. (2019a, 2019b). Example test sentences for each structure (including match and non-match items) along with their respective YouTube links are provided in Table 1.

Observation form

An observation form was developed to structure the collection of qualitative data on children's performance regarding a) their level of independence and type of tester support required and b) their level of engagement when interacting with the assessment. The form was informed by 1) previous TECS-E administrations, allowing us to anticipate potential responses and actions from participants, 2) existing literature on adult scaffolding of young children's technology use (Stock et al., 2004; Wood et al., 2016) and 3) levels of engagement as documented in Raspa et al. (2018). The form also allowed for tester reflections in open text.

Levels of independence during test taking were measured by the types of prompts requested or required by the child (and therefore provided by the tester), to complete the test. Prompts were categorized as physical, verbal and emotional-verbal. Physical prompts referred to any movement or gesture by the adult to support testing, including touching the screen on the child's behalf. Verbal prompts included any verbal comments by the adult related to testing, excluding praise. Emotional-verbal prompts included verbal prompts that had an emotional purpose, i.e., praise, encouragement.

Levels of engagement were rated on a 5-point Likert scale like that used in Raspa et al. (2018), which represents where: (1) participant refusing to interact with the assessment;

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

(2) limited engagement; (3) moderate engagement; (4) active engagement; and (5) overly engaged or difficult to disengage. Moderate engagement applied for example, if the participant was willing to focus and interact with the test appropriately but required some prompting by the tester. Off task behaviour was also documented. The aim was to put sufficient supports in place to achieve active engagement whilst also enabling maximum independence.

Procedure

Pilot administration of TECS-E was by a qualified speech and language therapist, or a psychology graduate trained in test administration. Testing took place in a quiet room in a preschool setting. The tablet was placed in front of the child and the test was administered with varying levels of support depending on the needs of the children. TECS-E administrations were video recorded to allow coders to document prompts and errors in accordance with the observation form. The specific purpose of each prompt within the predetermined categories was recorded / coded by both testers independently, and consensus agreed following discussion. Engagement was also rated independently by both testers and consensus agreed for all ratings. Following this, coders worked collaboratively to identify the initial adaptations and functional features required to facilitate pre-school children's self-directed administration of TECS-E. Adaptations were agreed based on the most frequent types of prompting required from testers and their general observations on the barriers to children's ability to comprehend the test instructions and independently use the test. Adaptations were also informed by a review of the literature surrounding design features in app development, and existing educational apps for children (e.g., Things that go together, Dino Island, 123 Toddler, Reading Eggs and My house).

Results and discussion

Levels of independence / prompts and engagement

Most physical prompts required related to touching the screen in some way – either to demonstrate a swipe (to move on to the next item); play or replay an item; or input the child's response. Difficulty with a swipe movement for young children has been noted in previous literature (Brooks, 2012) as well as young children's lack of dexterity to target small icons (Hanna et al., 1999). Verbal prompts included step by step instructions (to move on to the next item); encouragement to the child to respond; responses to direct requests for support; providing affirmation; redirection to the task; and maintaining the child's attention. Verbal-emotional prompting was used to ensure the child remained motivated/ engaged with more frequent praise and redirection needed at transition points (e.g., before and after reward stars) and towards the end of the assessment. For a full list of visual, verbal and emotional-verbal prompts children required when completing the initial iteration of the test see Table S1 in supplemental material.

In relation to engagement, 4 children were rated as actively engaged, 2 moderately and 1 had limited engagement. Factors which negatively affected children's engagement included the length of the testing period, frequency of technology failures and difficulty using the technology independently. Testers observed that children's off task behaviours increased as they got closer to the end of the test. Difficulties maintaining attention when 'loading' or following technological errors have been noted in previous literature (Rust et al., 2014).

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

Adaptations required

TECS-E adaptations required were categorised under three main subtypes: those which aimed to facilitate comprehension of test instructions; those which facilitate independent technology use; and those aimed at promoting attention and increased engagement. The adaptations made to TECS-E as well as their basis in the observational data are outlined in Table S2 in supplemental material.

Facilitating comprehension of test instructions

A number of key issues arose in relation to the comprehension of test instructions. Firstly, children often appeared to ignore the audio content and seemed happy to base judgements on what occurred in the animations alone. This observation was reinforced when on a couple of occasions the audio failed but children continued to make judgements without noticing that the audio was missing. The resulting adaptation was that in the TECS-E app, audio is spoken by an animated character rather than a 'disembodied' voice accompanying each animation. In addition, instructions on how to use the app are given by the character who points to different parts of the screen, therefore increasing the saliency of the instruction. Hiniker et al. (2015) have found that children between 3 and 3;06 can successfully follow in-app audio instructions and on-screen demonstrations. Our use of an animated character is in keeping with Hanna et al. (1999), who suggest that a 'character' giving directions is useful to direct children's attention. Moreover, McKnight and Fitton (2010) highlight that instructions are better understood when visuals and audio are presented together. Secondly, children made moral rather than truth value judgements on the animated content. Some children qualified their responses with comments such as that present wasn't hers" or "he shouldn't be climbing trees". This error was thought to originate from testers asking "did the lady get it right" (when the audio and animation were congruent) or did she get it wrong (where the animation and audio were incongruent) during test administration. The resulting adaptation was that the word "wrong" was avoided in the final version of the test and the phrase "made a mistake" was used. Thirdly, children were observed to base some responses on the emotions reflected in the animated content. We believe this was influenced by the response button symbols which were represented by a smiley and sad face. For example, when someone in the animation was crying, children tended to select the sad face. Consequently, we changed the response buttons to a smiley and 'oops' face (to reflect a mistake). Finally, the truth value judgement concept was difficult for children to grasp at this age and the practice items were not sufficient to ensure they knew what to do. To support this, a training game was designed to be administered prior to the test items to teach children the process and the truth value judgment concept. The game progressed from asking children to simply "touch here" to asking them to make their own judgements on the answer and press the corresponding button without prompting. Training progressed from making judgements on individual objects (such as cup, toothbrush) to simple sentences "The boy is eating" and finally complex sentences similar to those in the main test e.g., "She dried the boy that was sitting."

Facilitating independent technology use

An analysis of physical prompts required by children allowed us to determine functional features that would facilitate more independent engagement with the test. The most

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

frequent physical prompt required was touching the screen on a child's behalf to play the next test item or because children had difficulty carrying out more complex touch gestures such as swiping. This slowed the pace of testing and reduced children's independence. The final version of the test has been designed to eliminate the need for these physical prompts by automatically progressing to the next item once a response is selected. In addition, the only touch gesture now required is a single tap and buttons are designed to be large and spaced apart to minimise errors. This is keeping with Brooks (2012) and McKnight and Fitton (2010) who noted single taps to be the most intuitive gestures for children and Anthony et al. (2014) who suggested that children lack dexterity to target small icons and often miss and tap the 'gutter'.

Maintaining attention and increasing engagement

Many children had difficulty attending to the full test without frequent praise and redirection, particularly as the assessment progressed. Some also required breaks during the administration. Consequently, the test was shortened, such that two types of each structure were included (subject and object relatives, *think* and *know* sentential complements and *before* and *after* adverbials). In addition reward 'star' animations were incorporated more frequently. The star animations included templates/shadows to indicate all the stars that could be collected and to allow children to see how close they were to the end of the test (conveying similar information to a progress bar). This is supported by Hiniker et al. (2015) who found that progress bars support engagement of children by 4 years old. In addition, if children didn't respond, audio prompts were provided as a reminder within a prescribed time frame. This adaptation was aimed at redirecting children back to the task if engaged in off-task behaviour.

Following implementation of all adaptations, TECS-E was developed as an IOS app. Figure 1 shows images of the new interface – on the top is the image for the sentence *If the boy was taller he could reach the teddy*, along with the pause, play, oops and item correct buttons, underneath shows the image of Bella giving instructions.

Main study

Method

Study Design

The main study was observational. A within-subjects design was used, whereby participants completed the TECS-E app twice (repeated-measures) to establish test-retest reliability. It was not possible to blind participants or researchers to the nature of the test when it was being taken (independent vs supported); however, the final analysis of the data was conducted blind to test-order.

Power analysis

Our ultimate goal is to use the TECS-E as a self-directed independent measure of children's understanding of complex sentences and to reveal individual differences in children's performance on the test from as young an age as possible. Therefore, we aimed to optimise the self-directed version of the test to achieve observed reliability of at least .75 with supported administration of the same test. While previous studies led us to expect a

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

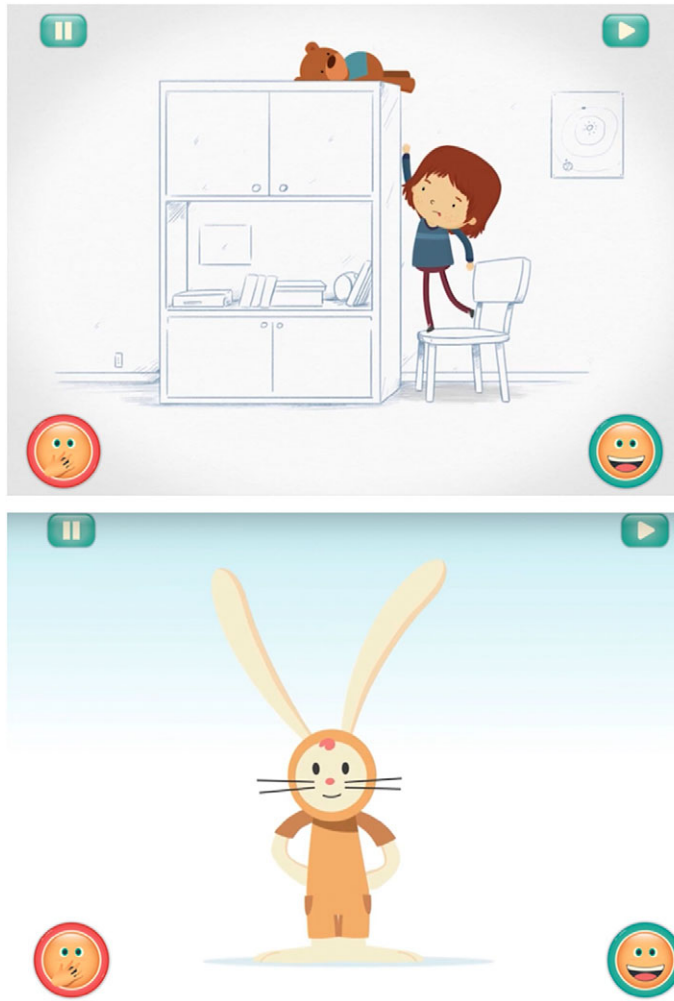

Figure 1. Adapted TECS-E Interface

test-retest correlation of around 0.75 (or better), we were concerned that this level of performance might not be achieved in an age group as young as 4 to 4;06 years. Thus, the sample size was based on our aim to confidently identify a poorly performing test, where the test-retest r is < 0.75. Our sample size was therefore 112 participants (paired observations), based on a one-sided, dividing hypothesis test of H0: r >= 0.75 vs H1: r < 0.75 with a 10% type 1 error (concluding r < 0.75 when it isn't), based on a normal model following Fisher's transformation of r. This results in 95% power to correctly reject H0 when the true r < 0.60. Our rationale for increasing the acceptable type 1 error, and thus reducing the type 2 error, is that incorrectly accepting a poorly performing test (i.e., making a type 2 error) is a more dangerous error than incorrectly rejecting a test that is performing adequately (a type 1 error), since the latter error would likely lead to continued testing and development, whereas the former could lead to changes in practice.

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

Stopping rule

We anticipated that the test-retest r might be so poor in the 4 to 4;06 year olds that we could confidently detect poor performance before the full sample was recruited. If that was the case, then it would be advantageous to reject the use of the test in that age group as early as possible, and start recruiting older participants, to try to identify the youngest age group for which the self-directed test could be used. Consequently, we used a group-sequential testing procedure. This means we planned 3 interim one-sided hypothesis tests of H0: r >= 0.75 vs H1: r < 0.75 when n is equal to 28, 56, and 84 participants, respectively. Because, the potential for early stopping can needlessly inflate the overall type 1 error rate unless we appropriately modify the rejection rules (i.e., the p-values that would lead us to reject the null), we used O'Brien-Fleming (O'Brien & Fleming, 1979) boundaries to adjust the p-values across the 4 potential tests (the three interims tests plus the final one) to 0.0024, 0.023, 0.0522, and 0.0797 (rather than naively rejecting the null when $p < 0.10$ at any point across the series of tests). This meant that we required substantially more evidence to reject the null for the interim hypothesis tests, as a function of how early the interim test was conducted. Rejection of H0 r >= 0.75 at any of these tests (e.g., a p-value lower than the above listed values) would lead us to start recruiting in the next older age group of 4;06 to 5 year olds for the remaining duration of the study.

Participants

Twenty-two typically developing children aged 4–4;05 years were initially recruited. This initial age range, which is a stage of continuing growth in complex syntactic language development, was chosen based on previous work carried out by Frizelle et al. (2017b, 2019a, 2019b). It was followed by three subsequent recruitment drives, children who were aged 4;06 to 4;08 ($n = 34$); 4;09 to 4;11 ($n = 22$) and 5 to 5;05 ($n = 113$). The recruitment procedure, participant inclusion criteria and areas of pre/school involvement were as outlined in the pilot study. For the final group we made the decision to recruit the full sample size (without interim analysis). We hypothesized better agreement for this age group and did not have capacity to continue recruitment beyond that sample. As previously outlined the Ling six sound test was used to screen children's hearing. Demographic information on study participants is shown in Table 2.

Randomization

At study entry, participants were randomly allocated to one of two groups (independent first vs supported first). Restricted randomization lists were prepared using a statistical software random assignment generator. The randomization lists and resulting allocations were prepared by the second author.

Variables

In this test-retest design, the within-subject independent variable of interest was Session (1 or 2). The measured variable was the total score on the adapted Test of Complex syntax-Electronic (TECS-E) web-based app. In addition to the adaptations outlined in Table S1, for the purposes of establishing the feasibility of self-directed administration we administered a shorter version of TECS-E. The test included 48 animations depicting 2 types of relative clause, 2 types of adverbial clause and 2 types of sentential complement (each structure represented 8 times). Catch items and star animations were also incorporated.

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

Table 2. – Demographic information on main study participants per age group

Participant Age Group	N	Mean Age (in months)	Age SD	Males	Females	SES*
4;00 to 4;05	22	50.86	1.96	12	10	40.91% - very disadvantaged 45.45% - marginally below average 13.64% - affluent
4;06 to 4;08	34	55.12	0.77	13	21	14.71% - marginally below average 58.82% - marginally above average 26.47% - affluent
4;09 to 4;11	22	57.95	0.90	11	11	72.73% - marginally above average 27.27% - affluent
5;00 to 5;05	113	62.85	1.77	48	65	0.88% - disadvantaged 4.43% - marginally below average 51.33% - marginally above average 43.36% - affluent

*SES was assessed at neighbourhood level by identifying the location of the participants' schools and pre-schools on the Pobal deprivation indices map. This is a free geographical information system, run on behalf of the Irish Government, which profiles deprivation under the following categories: extremely affluent, very affluent, affluent, marginally above average, marginally below average, disadvantaged, very disadvantaged, and extremely disadvantaged.

Procedure

All participants were assigned an identification code to facilitate cross-referencing between the supported and independent assessments. Supported assessments were administered by one of three qualified speech and language therapists or a postdoctoral researcher who was trained in test administration. Prior to beginning testing, a series of instructions were agreed between each tester as to what was permitted within each testing condition. Following the integration of identified supports to allow for self-directed testing, the aim was that the supported and self-directed assessments would be delivered with identical instructions and in as similar a manner as possible. However, although we tried to keep differences to a minimum, we did anticipate some differences. In line with good clinical practice, we expected that the supported delivery would involve the tester establishing rapport with the child for a short period before beginning the assessment. It was also agreed that if a child required clarification regarding a particular part of the assessment process, the tester would be permitted to answer. In addition, if a child appeared hesitant testers were permitted to give encouragement. When testing in the supported condition testers were asked to log this information. In the case of a particularly sociable child, it was acknowledged that they may engage with the tester throughout the assessment, thereby establishing a level of affinity and reducing the stress that accompanies a 'testing' experience. On the other hand, it could be argued that a child completing the assessment independently in pre/school, without the involvement of an unfamiliar adult could feel less pressure to perform. Those administering the supported test were asked to confine themselves to the scripted instructions as much as possible, without jeopardizing their relationship with the child.

Both supported and self-directed testing took place in a quiet room in the pre-school or school that the children attended. A tablet was placed in front of the child and the test was administered in the agreed standardized way, with or without the support of the tester. In the independent condition testers were in the room with the child but completing their own work at a nearby desk. Participants completed the test twice and were randomly

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

assigned to two subgroups for counterbalancing. One group received the supported test first followed by the independent assessment whereas the other group completed the independent assessment first, followed by the supported administration. The second testing session for each participant took place within 3 weeks of the first one but never on the same day. This timeframe was chosen to limit the influence of developmental changes on the completion of the second test, while at the same time reducing practice effects by not administering the tests too close together.

Data analysis

All statistical analyses were performed using R Statistical Software (R Core Team, 2018). Code for replicating all pre-specified and any additional reported analyses are available via the OSF (https://osf.io/emfcy/). Test – retest reliability, between the supported and independent use of the modified TECS-E, was examined using a Pearson's product moment correlation coefficient for paired samples. Agreement between the two test scores was evaluated visually using a Bland-Altman plot, and the mean difference in scores and their 95% limits of agreement (based on the sample SD of the differences) was calculated. To further evaluate the agreement between the two tests, we estimated a series of multilevel models, where the two test scores were nested within children. An empty model with no covariates was used to estimate the intra-class correlation, where we expected relatively small within-subject variation relative to between-subject variation. Then we added an indicator variable for test version, to estimate the mean difference in test scores; following this we added age (days) and an interaction between age and test version as covariates, to evaluate whether this aspect of agreement varied across ages. The usefulness of this interaction term was evaluated with a likelihood ratio test for nested models. Estimates are reported below alongside frequentist 95% confidence intervals and exact p-values.

Data exclusions

Data from the TECS-E app were collected and stored by Gorilla (https://gorilla.sc/), a cloud-based tool for collecting and storing data in the behavioural sciences. It was intended that data would be reported for all who met the study entry criteria and consented into the study. However, there were IT difficulties and data from 28 participants were lost, 9 of which were from the 5 to 5;05 age group. An additional six participants did not complete either of the TECS-E tasks within the required time frame (due to being absent from the pre/school) and were consequently excluded. However, each of the excluded participants were replaced, so that a total of 113 complete datasets were still collected. One additional participant was excluded at the analysis stage for having a test score >5 SD from the sample mean, resulting in an analysis set of 112 children.

Deviations from pre-registration

We had stated in our pre-registration that we would use GAMLSS (generalized additive model for location, scale, and shape) to model each child's difference in test scores as a function of their age, allowing for heteroskedastic errors across ages. However, there was no age effect on the mean for children between 5 – 5;05 years and no evidence of heteroskedasticity by age; therefore, we did not do this. We had also planned to apply a Box-Cox transformation to scores prior to analyses, and where relevant, to back-transform the results

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

to the original score metric. However, the transformation did not make any appreciable difference to the observed outcome distribution and therefore we report here on data that have not been transformed. We have included a Figure in our supplemental material showing the distribution of total trial items with and without Box-Cox transformation.

Results

Our first hypothesis was that the re-test reliability in the accuracy scores between the independent and the examiner supported administration of TECS-E might be poor for the children between 4 and 4;05 years. Following supported and independent completion of the test with 22 children in this age range, we made the decision to stop testing. Qualitative observations highlighted several difficulties for children at this age and indicated a level of guessing in both supported and independent conditions.

Qualitative observations

Maturity / sustained attention

For the younger children in this age group there was considerable variability in their readiness to take part in a task that involved completing an app. This appeared to be less problematic for children closer to 4;05 years. The younger children sought a lot of positive reinforcement and encouragement from the testers in the initial stages. In contrast other children were so motivated to interact with the app, that they did not listen to initial instructions from the tester and were busy pressing a button to select an answer before the tester had a chance to give guidance. Children who completed the task in a supported context first appeared to expect to receive the same level of support for the subsequent independent version. In these cases, it was challenging for the research assistant to explain to young children that they needed to try to complete the task independently. Some of the younger children were also distracted by the animation itself and did not appear to simultaneously process the verbal information. For instance, one child in the independent context commented on the animation to the tester and related it to his own life (e.g., "My brother broke a window in my house"), rather than remembering to complete the truth value judgment task. Overall, children required frequent praise, breaks, and redirection to stay on task in both the supported and self-directed test conditions, particularly as the assessment progressed.

Understanding the truth-value judgement task

The design of the practice section was such that the demands placed on the child appeared to increase too quickly. Following the presentation of 6 training items focusing on a common object (e.g., This is a ball), children were only presented with 2 simple action sentences e.g., "The girl is wearing glasses." if they got some of the training items wrong. If all the training items were correct children progressed straight onto the complex items. However, while children seemed to have some understanding of the truth value judgment task when presented with common objects (e.g., This is a ball) they appeared to forget the concept when presented with complex sentences and reverted to indicating *yes* for all items or would move their finger back and forth between buttons before randomly selecting an answer. For example, when presented with an incongruent sentence–animation pair, one

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

child was recorded successfully determining that what she had seen did not match what she had heard, but had difficulty completing the task, saying "No she didn't. But does that mean the bunny is right or not?". When children did not provide the correct answer for the complex practice items, this triggered an automatic response "Uh-oh, that's the wrong button!". While some children appeared to learn from this feedback, others looked embarrassed and were recorded saying "*I'm not good at this game*", "*This is too hard*".

Use of the buttons and pace

Some children found it difficult to remember the meaning of the answer buttons. For example they provided the correct answer verbally "*No, she's riding a bike*", but selected the wrong button. Other children became fixated on the replay and pause buttons, which they had the autonomy to press at any stage. This made the test longer, leading to increased test fatigue, and decreased co-operation. Overall, it seemed that the pace of the app was too fast for this young age group. For instance, if a child did not respond within a few seconds, the app prompted the child to answer "Did you see that video? If you want to watch it again press here". In these instances, some children became distracted and fascinated by the replay feature and needed to be reminded to continue with the app.

Consistent with our pre-registration, we then examined test-retest reliability with children aged 4;06–4;11 months (n = 55)[2]. We did not carry out an interim analysis at n= 28 as there was significant variability in children's responses at this age and we wanted to ensure a reasonable sample to inform our findings. The intra-class correlation (ICC) between total scores on the supported versus independent testing was .57 (p = .0103) and based on an adjusted p value of .023, indicated that we should reject the use of the test in that age group. This was supported by our qualitative observations.

We examined test-retest reliability in scores for children aged 5- 5;6 years based on our full sample (n= 112; Figure 2). The estimated paired-sample Pearson's correlation (for all items) was 0.71 (95% CI 0.6 to 0.79; p = 0.17 for H0: r >= 0.75 vs H1: r < 0.75).

An estimated paired-sample Pearson's correlation was also completed for the congruent and incongruent items separately. The correlation was similar across both sets of items (r = .76, 95% CI 0.67 to 0.83 for the congruent items), (r = .74, 95% CI 0.64 to 0.81 for those that were incongruent). Scatter plots for these analyses are available in supplemental materials (Figures S2 and S3).

We also evaluated agreement in scores comparing the supported and independent test conditions using a Bland–Altman analysis (Bland & Altman, 1995). The mean difference was -2.27 and the 95% limits of agreement were -10.8 to 6.3. Visual inspection of the Bland-Altman plot did not reveal any concerning patterns or trends (Figure 3). The above results were confirmed using multi-level models (Table 3) where the estimated mean difference in scores was -2.27 (95% CI -3.08 to -1.46) in the Bias model, with an ICC of 0.71. The addition of age or an age by test interaction did not appreciably improve model fit.

In our second hypothesis we predicted that the same rank ordering of constructions would emerge between the independent and supported test administrations and overall, this was borne out in our data (see Figure 4). A rank order was assigned to each of the constructions for mean items correct in the independent and supported use of the test, and a Spearman rank order correlation was calculated (r = .94).

[2]Note data were collected for 56 children but one outlier was removed before the analysis.

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

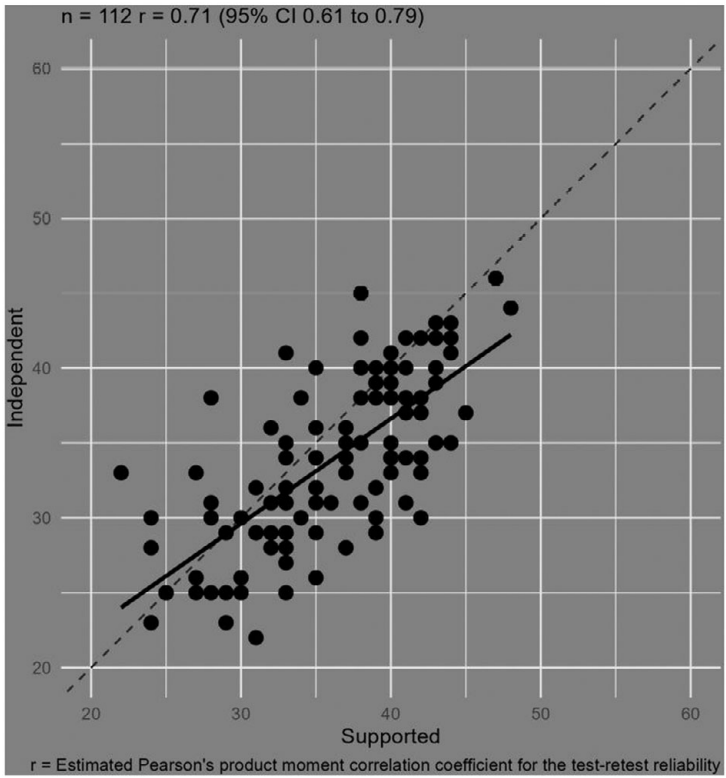

Figure 2. Scatter plot of total trial items correct in the supported and independent test conditions.

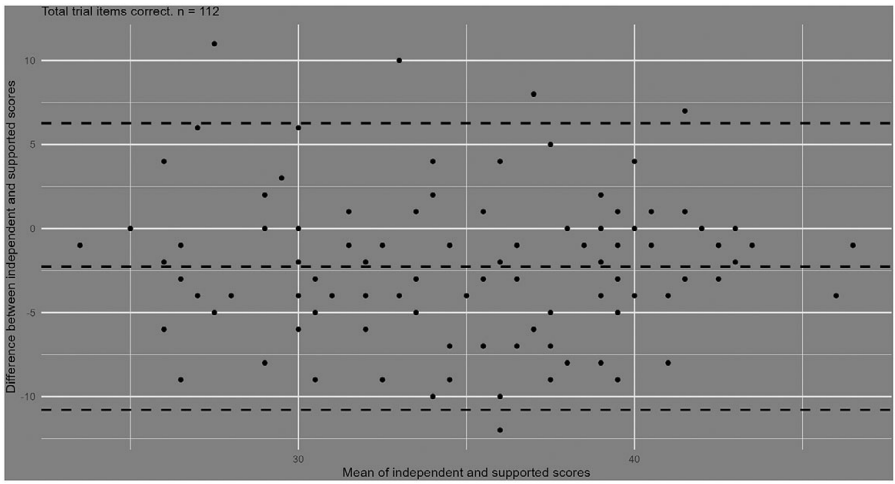

Figure 3. Bland-Altman plot for agreement in total trial items correct comparing the supported and independent test conditions.

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

Table 3. Multilevel models of total trial items correct

	Empty			Bias			+Age			+Age interaction		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
Intercept	35.13	34.14 – 36.13	**<0.001**	36.27	35.20 – 37.34	**<0.001**	36.27	35.19 – 37.34	**<0.001**	36.27	35.19 – 37.34	**<0.001**
Test condition (independent vs supported)				−2.27	−3.08 – −1.46	**<0.001**	−2.27	−3.08 – −1.46	**<0.001**	−2.27	−3.08 – −1.46	**<0.001**
Age (months)							0.01	−0.54 – 0.57	0.960	0.13	−0.47 – 0.73	0.676
Age by Test interaction										−0.23	−0.68 –0.23	0.325
Random Effects												
σ^2	11.95			9.46			9.46			9.46		
τ_{00}	22.45 $_{id}$			23.69 $_{id}$			23.95 $_{id}$			23.95 $_{id}$		
ICC	0.65			0.71			0.72			0.72		
N	112 $_{id}$			112 $_{id}$			112 $_{id}$			112 $_{id}$		
Observations	224			224			224			224		
Marginal R^2 / Conditional R^2	0.000 / 0.653			0.037 / 0.725			0.037 / 0.727			0.038 / 0.728		

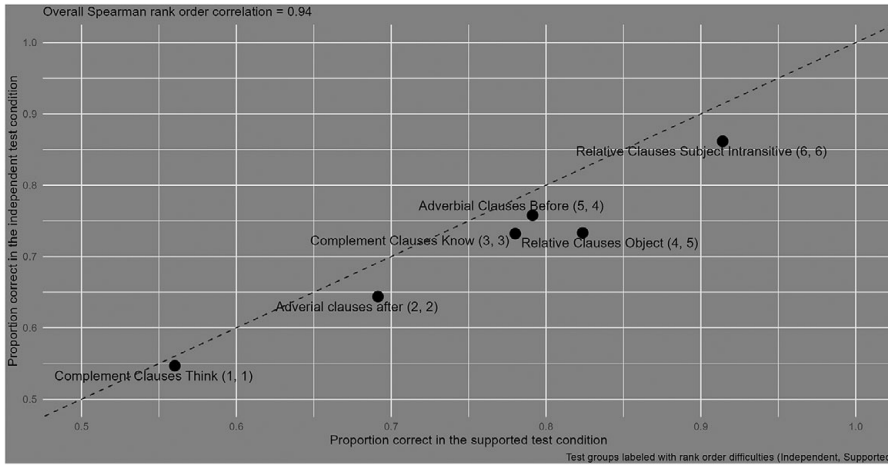

Figure 4. Proportion of correct answers across all participants (n = 112) in each test group, comparing the supported and independent test conditions.

The plot shows each item label with two pairs of numbers, one indicating the rank order of item difficulty in the independent condition and the other in the supported condition. Number pairs that are the same indicate the same ranking in relation to order of construction difficulty. The dashed line indicates where points would fall if the difficulty of a construction type was the same in both conditions, and points that fall below the line indicate greater difficulty in the independent test condition. The plot shows that complement clauses '*think*' were the most difficult in both conditions, followed by adverbial clauses '*after*' and complement clauses *know*. In addition, intransitive subject relative clauses were the easiest in both conditions and this is in line with previous findings, when comparing different types of relative clauses (see Frizelle et al., 2017b, 2019a). The only difference in rank ordering between both conditions was adverbial clauses *before* and object relative clauses. In the independent condition object relatives were ranked 4th versus a ranking of 5th in the supported condition whereas the reverse was the case for adverbial clauses *before*. However children's performance overall was very similar on these constructions and the proportion of children who performed better on one family of constructions versus the other was small (adverbial *before* .79 vs .76 and object relatives .82 vs. 73).

Discussion

Our main study aimed to investigate the level of agreement between two methods of administering the TECS-E language assessment app – one 'supported' (similar to what would happen in a clinical setting) and one self-directed (more aligned to online methods of testing) and to establish how young the self-directed method could be used reliably. In addition, we aimed to ascertain if the two methods would agree in terms of the order of difficulty of the specific constructions, further validating the use of the independent self-directed method of assessment as a reliable indication of children's understanding of complex sentences. Consistent with our hypothesis, our results revealed that children between 4 and 4;05 years had difficulty completing the test as a self-directed app. Furthermore, the app posed problems for children at this age even when administered

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

in a more supported context. Although less problematic, difficulties also emerged for children between the ages of 4;06 and 4;11 and our statistical analyses suggested that the test was not reliable when administered as a self-directed assessment for this age group. Finally, we examined test retest reliability for children aged 5 – 5;05 years. Although there is considerable variability in children's performance at this age, our findings suggest that the TECS-E self-directed app is a reliable method to assess children's understanding of complex sentences at this age. This is further validated by our finding that both independent and supported use revealed a similar order of construction difficulty.

Qualitative observations with younger children

The finding that many children between 4 and 4;11 had difficulties using the modified app, whether supported or completed independently, warrants further discussion. While previous studies reporting on the use of the TECS-E with younger children (see Frizelle et al., 2019a, 2019b) presented the test items on computer (with pre-recorded audio), the therapist controlled the pace of the test; could scaffold the child through the practice items; could repeat test items live when children got distracted; and could assist the child to press the correct button if they gave the right answer orally. However, because much of the TECS-E app was designed to be automated, the amount of therapist support that could be given, even in the supported context, was quite limited. We had designed the app with auto-progression to the next item aiming to increase independence and make the process less cumbersome and therefore more enjoyable. However, in-built instructions and automatic progression (following the child's response) meant that the therapist could not easily scaffold the child's learning during the practice items or prompt during the test itself. Consequently, many children did not provide the correct answer for the complex practice items. Because this generated an automated response from 'Bella' '*Uh-oh, that's the wrong button*' many of the younger children were embarrassed and appeared to lose confidence. Within the main test items, if the tester attempted to prompt children in a natural way, the prompt from Bella would also play and result in the character and the tester speaking at the same time. We did not have a 'back' button as children tend not to use this feature (Nielsen, 2002). Although there was a replay button this could not be used once a choice was made. In addition, the pause button, when pressed, triggered a black screensaver which obscured the image of the response buttons, meaning that the therapist could not use the visuals of the buttons to reinforce the test instructions during the practice items.

The truth-value judgement concept of the test was difficult to grasp for the younger children in this age range, particularly as language complexity increased. It is possible that the competing demands of learning the task as well as processing complex test items placed too great a load on working memory for these young children. The task required that children store the sentence they heard with the corresponding video; decide if what they heard matched what they had seen; and recall which button they needed to press based on their response. Our observations suggest that in the training game there were not enough practice items at the simple sentence level for children to consolidate the truth value judgement concept while simultaneously remembering which button to press. This was particularly evident in responses where children gave the correct response verbally but then pressed the wrong button. Further testing outside of the context of this reliability study revealed that very young children responded better when, following the presentation of the pre-recorded test item, the therapist repeated the sentence (e.g., *He found the girl that was hiding*) and asked the child 'yes' or 'no'?. The therapist then pressed the appropriate button according to the child's response.

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

Test/re-test reliability 5–5;05 years

Our finding that children between 5 and 5;05 years could reliably complete the TECS-E as a self-directed app was in line with our hypothesis. While we had anticipated a slightly higher ICC of ≥.75, applying the 95% confidence interval, our point estimate of .71 is not inconsistent with this level of reliability (95% CI .61– .79). Generally ICC values of .60 or .70 and higher are considered acceptable criterion levels of test-retest reliability (Anastasi, 1998; Ruano et al., 2016). We can therefore conclude that when completed independently as a self-directed app, the TECS-E in its current format is a reliable method of assessing children's understanding of complex sentences from 5 years. Qualitatively, most children in this group presented as confident and capable of completing the task. While study order was controlled for in our analysis, from observation it appeared that children who completed the supported version first found it easier to complete the test independently and this was reinforced by our findings that children scored 2.3 points lower on the independent compared to the supported administration. Overall, our findings are in keeping with Csapó et al. (2014) who, despite finding a high degree of correlation between face to face and online test administrations, found children's performance was lower on computer based tests than on face to face equivalents. In contrast to our younger group who either sought or became reliant on support (depending on which version was administered first), for the most part, children at this age didn't seek support to complete the assessment. For those who did, following minimal support in the very early stages, they quickly understood the concept of the task. We do not have comparator studies where a truth value judgement task was used, or where independent versus supported computerized testing were compared.

Finally, the use of TECS-E as a self-directed language assessment with children over 5 years is further validated by our finding that overall, both independent and supported use revealed a similar order of construction difficulty. While two construction types were reversed in ranking between both conditions (4^{th} versus 5^{th} and 5^{th} versus 4^{th}) the proportion of children who were correct across these constructions was very close. A similar construction hierarchy between conditions was in keeping with our hypothesis which was informed by previous work comparing two different methods of assessing young children's understanding of different types of relative clauses (sentence repetition and multiple-choice sentence picture matching tasks) (Frizelle et al., 2017).

Changes to the current format

Having administered the TECS-E app in both supported and independent contexts with 97 children between 4 and 4;11 years (19 for whom the data were lost), there are a number of changes that we believe would facilitate more reliable use of the test with children in this age group. These changes are also likely to be beneficial for children over 5 years and in particular for children at risk of language disorder or for those with DLD, for whom the test is ultimately designed. While we had endeavoured not to make the practice section too long (including only 2 simple sentences if children got a training item wrong), we plan to replace 2 of the training items with 2 simple sentences, and these will be presented regardless of children's performance on the training items. This should facilitate children's consolidation of the truth value judgement concept before the items become too linguistically challenging. When children respond incorrectly to the practice items we will no longer give the feedback "Uh-oh, that's the wrong button" and will instead use the phrase "Oops, that time I made a mistake" so that children will not lose confidence at an early stage in the assessment process. We will also slow the pace of the

https://doi.org/10.1017/S0305000923000545 Published online by Cambridge University Press

app so that children have a greater amount of time to respond before being prompted by Bella to make their choice. This includes preventing the screen from changing too quickly following receipt of a star. Children typically enjoyed counting the stars that they got and how many they still needed to get, to complete the app. However, in its current form the screen changes too quickly for some children to finish counting their stars. In animations where children are required to pay more visual attention to a particular detail to understand the sentence, we plan to make these more salient by adding discrete sound effects. This is consistent with Brooks (2012) who found that the addition of sound effects increases engagement. We will no longer use emojis (that convey an emotion) to represent the response buttons and instead will use a $\surd$ and X to indicate *yes* and *no* respectively. In addition, when the pause button is pressed, we will make the screen freeze so that the still of the animation and the buttons can still be seen, rather than the black screen that currently shows. Finally, we plan to convert the app from one that can only be used on an IOS platform, to a web-based app that can be downloaded and used across all platforms.

Limitations

Overall, it would have been preferable to pilot our initial set of adaptations with a larger number of children. However, the onset of Covid-19 meant that we could not access any more children at that stage in the project. To complete the work in a timely manner and within the schedule of the IT company, we had to progress with creating the app based on the feedback of this initial cohort. However, our work assessing 97 children between 4 and 4;11 has provided us with rich information to inform the final version of the tool to be used for norm referencing and standardization. It would also have been interesting to gather information on children's familiarity with apps across each age, as increased familiarity could potentially enhance children's ability to complete the app online.

Conclusion

This study shows that children between the ages of 4 and 4;06 years were not sufficiently scaffolded to complete the TECS-E app reliably (in its current form) in either a self-directed or supported context. Children between 4;06 and 4;11 years understood the process more reliably but the support of an adult yielded higher scores, reflecting a better understanding of language. Consequently, the independent versus supported contexts were poorly correlated. Finally, our results show that from 5 years of age TECS-E, when used as a self-directed app, is a reliable method to assess children's understanding of complex sentences.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0305000923000545.

Funding statement. This work was funded by a Lead Investigator award (ILP-POR-2019-003) from the Health Research Board, Ireland.

References

Anastasi, A. (1998). *Psychological testing* (6th ed.). New York: Macmillan.

Anthony, L., **Brown, Q.**, **Tate, B.**, **Nias, J.**, **Brewer, R.**, & **Irwin, G.** (2014). Designing smarter touch-based interfaces for educational contexts. *Personal and Ubiquitous Computing*, **18**(6),1471-1483.

Barak, A., & **English, N.** (2002). Prospects and limitations of psychological testing on the internet. *Journal of Technology in Human Services*, **19**(2-3), 65–89. http://doi.org/10.1300/J017v19n02_06

Bartram, D. (2006). Testing on the internet: issues, challenges and opportunitiesin the field of occupational assessment. In Bartram, D., Hambelton, R.K. (Eds.) *Computer-Based Testing and the Internet: Issues and Advances*. John Wiley & Sons, Ltd. pp. 13–37.

Bauer, R. M., **Iverson, G. L.**, **Cernich, A. N.**, **Binder, L. M.**, **Ruff, R. M.**, & **Naugle, R. I.** (2012). Computerized neuropsychological assessment devices: Joint Position Paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *Archives of Clinical Neuropsychology*, **27**(3), 362–373. http://doi.org/10.1093/arclin/acs027

Betz, S. K., **Eickhoff, J. R.**, & **Sullivan, S. F.** (2013). Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. *Language, Speech and Hearing Services in Schools*, **44**(2), 133–146. http://doi.org/10.1044/0161-1461(2012/12-0093)

Bilder, R. M. (2011). Neuropsychology 3.0: Evidence-based science and practice. *Journal of the International Neuropsychological Society*, **17**(1), 7–13. http://doi.org/10.1017/S1355617710001396

Birnbaum, M. H. (2004). Human research and data collection via the internet. *Annual Review of Psychology*, **55**, 803–832. doi:10.1146/annurev.psych.55.090902.141601

Bland, J. M. & **Altman, D. G.** (1995). Comparing methods of measurement: why plotting difference against standard method is misleading. *The Lancet*, **346**(8982), 1085–1087. https://doi.org/10.1016/s0140-6736(95)91748-984-9

Brooks, M. (2012). Best practices: Designing touch tablet experiences for preschoolers. Retrieved from http://www.sesameworkshop.org/wp_install/wp-content/uploads/2013/04/Best-Practices-Document-11-26-12.pdf.

Buchanan, T. (2003). Internet-based questionnaire assessment: appropriate use in clinical contexts. *Cognitive Behaviour Therapy*, **32**(3), 100–109.

Buchanan, T., **Heffernan, T. M.**, **Parrott, A. C.**, **Ling, J.**, **Rodgers, J.**, & **Scholey, A. B.** (2010). A short self-report measure of problems with executive function suitable for administration via the Internet. *Behavior Research Methods*, **42**(3), 709–714.

Carson, K., **Gillon, G.**, & **Boustead, T.** (2011). Computer-administrated versus paper-based assessment of school-entry phonological awareness ability. *Asia Pacific Journal of Speech, Language and Hearing*, **14**, 85–101.

Chan, J., **Adlof, S. M.**, **Duff, D.**, **Mitchell, A.**, **Ragunathan, M.**, & **Ehrhorn, A. M.** (2022). Examining the associations between parent concerns and school-age children's language and reading abilities: a comparison of samples recruited for in-school versus online participation. *Language, Speech, and Hearing Services in Schools*, **53**(2), 431–444. https://doi.org/10.1044/2021_lshss-21-00080

Csapó, B., **Molnár, G.**, & **Nagy, J.** (2014). Computer-based assessment of school readiness and early reasoning. *Journal of Educational Psychology*, **106**(3), 639–650. http://doi.org/10.1037/a0035756

Denman, D., **Speyer, R.**, **Munro, N.**, **Pearce, W. M.**, **Chen, Y.-W.**, & **Cordier, R.** (2017). Psychometric properties of language assessments for children aged 4–12 years: A systematic review. *Frontiers in Psychology*, **8**, 207–228. https://doi.org/10.3389/fpsyg.2017.01515

Diessel, H. (2004). *The acquisition of complex sentences*. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/cbo9780511486531

Diessel, H., & **Tomasello, M.** (2000). The development of relative clauses in spontaneous child speech. *Cognitive Linguistics*, **11**, 131–151. https://doi.org/10.1515/cogl.2001.006

Diessel, H., & **Tomasello, M.** (2005). A new look at the acquisition of relative clauses. *Language*, **81**, 1–25. https://doi.org/10.1353/lan.2005.0169

Feenstra, H. E. M., **Murre, J. M. J.**, **Vermeulen, I. E.**, **Kieffer, J. M.**, & **Schagen, S. B.** (2018). Reliability and validity of a self-administered tool for online neuropsychological testing: The Amsterdam Cognition Scan. *Journal of Clinical and Experimental Neuropsychology*, **40**(3), 253–273. https://doi.org/10.1080/13803395.2017.1339017

Frizelle, P., **Harte, J.**, **Fletcher, P.**, & **Gibbon, F.** (2017a). Investigating the effect of regional native accents on sentence comprehension in children with language impairment. *International Journal of Speech-Language Pathology*, **22**, 1–13. http://doi.org/10.1080/17549507.2017.1293734

Frizelle, P., **O'Neill, C.**, & **Bishop, D. V. M.** (2017b). Assessing understanding of relative clauses: a comparison of multiple-choice comprehension versus sentence repetition. *Journal of Child Language*, **44**(6), 1–23. http://doi.org/10.1017/S0305000916000635

Frizelle, P., Thompson, P., Duta, M., & **Bishop, D. V. M.** (2019a). Assessing Children's Understanding of Complex Syntax: A Comparison of Two Methods. *Language Learning*, **69**(2), 255–291. http://doi.org/10.1111/lang.12332

Frizelle, P., Thompson, P. A., Duta, M., & **Bishop, D. V. M.** (2019b). The understanding of complex syntax in children with Down syndrome [version 2; referees: 3 approved] *Wellcome Open Research*, **3**:140 https://doi.org/10.12688/wellcomeopenres.14861.2

Frizelle, P. & **Fletcher, P.** (2014). Relative clause constructions in children with specific language impairment. *International Journal of Language & Communication Disorders*, **49**(2), 255–264. https://doi.org/10.1111/1460-6984.12070

Gallagher, A., Bridgeman, B., & **Cahalan, C.** (2000).The effect of computer-based tests on racial/ethnic, gender, and language groups (*GRE Board Professional Report No. 96-21P*). Princeton, NJ: Education Testing Service.

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F.,Chatterjee, G., & **Wilmer, J. B.** (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, **19**(5), 847–857. doi:10.3758/s13423-012-0296-9.

Gosling, S. D., Vazire, S., Srivastava, S., & **John, O. P.** (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist*, **59**, 93–104.

Hamann, C., Schultze-Lutter, F., & **Tarokh, L.** (2016). Web-based assessment of mental well-being in early adolescence: A reliability study. *Journal of Medical Internet Research*, **18**(6), e138–6. http://doi.org/10.2196/jmir.5482

Hanna, L., Risden, K., Czerwinski, M., & **Alexander, K. J.** (1999). The role of usability research in designing children's computer products. *The design of children's technology*, 3–26.

Haworth, C. M. A., Harlaar, N., Kovas, Y., Davis, O. S. P., Oliver, B. R., Hayiou-Thomas, M. E., Frances, J., Busfield, P., McMillan, A., Dale, P. S., & **Plomin, R.** (2007). Internet cognitive testing of large samples needed in genetic research. *Twin Research and Human Genetics*, **10**(4), 554–563. https://doi.org/10.1375/twin.10.4.554

Hiniker, A., Sobel, K., Hong, S. R., Suh, H., Kim, D., & **Kientz, J. A.** (2015, June).Touch screen prompts for preschoolers: designing developmentally appropriate techniques for teaching young children to perform gestures. In Proceedings of the *14th International Conference on Interaction Design and Children* (pp.109-118). ACM.

Hoffman, L. M., Loeb, D. F., Brandel, J., & **Gillam, R. B.** (2011). Concurrent and construct validity of oral language measures with school-age children with specific language impairment. *Journal of Speech Language and Hearing Research*, **54**(6), 1597–1608. http://doi.org/10.1044/1092-4388(2011/10-0213)

Kalff, A. C., De Sonneville, L. M., Hurks, P. P., Hendriksen, J. G., Kroes, M., Feron, F. J., Steyaert, J., van Zeben, T. M., Vles, J. S., & **Jolles, J.** (2005). Speed, speed variability, and accuracy of information processing in 5 to 6-year-old children at risk of ADHD. *Journal of the International Neuropsychological Society: JINS*, **11**(2), 173–183. https://doi.org/10.1017/s1355617705050216

Kidd, E., Brandt, S., Lieven, E. & **Tomasello, M.** (2007). Object relatives made easy: A cross-linguistic comparison of the constraints influencing young children's processing of relative clauses. *Language and Cognitive Processes*, **22**(6), 860–897. https://doi.org/10.1080/01690960601155284

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & **Couper, M.** (2004). Psychological research online: Report of Board of Scientific Affairs' Advisory Group on the conduct of research on the internet. *American Psychologist*, **59**(2), 105–117. http://doi.org/10.1037/0003-066X.59.2.105

Lo, C. H., Rosslund, A., Chai, J. H., Mayor, J., & **Kartushina, N.** (2021). Tablet assessment of word comprehension reveals coarse word representations in 18–20-month-old toddlers. *Infancy*, **26**(4), 596–616. https://doi.org/10.1111/infa.12401

Maguire, K. B., Knobel, M. L. M., Knobel, B. L., & **Sedlacek, L. G.** (1991). Computer-adapted PPVT-R: A comparison between standard and modified versions within an elementary school population. *Psychology in the Schools*, **28**, 199–205.

Marsh, J., Plowman, L., Yamada- Rice, D., Bishop, J., Lahmar, J., Scott, F., Davenport, A., Davis, S., French, K., Piras, M., Robinson, P., Thornhill, S., & **Winter, P.** (2015). *Exploring play and creativity in pre-schooler's use of apps: Final project report.* http://www.techandplay.org/reports/TAPFinalReport.pdf

McKnight, L., & **Fitton, D.** (2010, June).Touch-screen technology for children: giving the right instructions and getting the right responses. In Proceedings of *the 9th international conference on interaction design and children* (pp.238–241).ACM.

Naglieri, J. A., **Drasgow, F.**, **Schmit, M.**, **Handler, L.**, **Prifitera, A.**, **Margolis, A.**, & **Velasquez, R.** (2004). Psychological testing on the internet: new problems, old issues. *American Psychologist*, **59**(3), 150–162. http://doi.org/10.1037/0003-066X.59.3.150

Nielsen, J. (2002). Kids' corner: Website usability for children. Jakob Nielsen's Alertbox.

O'Brien, P. C., & **Fleming, T. R.** (1979). A multiple testing procedure for clinical trials. *Biometrics*, **35**(3), 549–556. http://doi.org/10.2307/2530245

Parsey, C. M., & **Schmitter-Edgecombe, M.** (2013). Applications of technology in neuropsychological assessment. *The Clinical Neuropsychologist*, **27**(8), 1328–1361. http://doi.org/10.1080/13854046.2013.834971

Paul, R., & **Norbury, C. F.** (2012). "Chapter 2: Assessment," in *Language Disorders from Infancy through Adolescence: Listening, Speaking, Reading, Writing and Communicating*, 4th Edn, eds **R. Paul** and **C. F. Norbury** (St Louis, MI: Mosby Elsevier), 22–60.

Raspa, M., **Fitzgerald, T.**, **Furberg, R. D.**, **Wylie, A.**, **Moultrie, R.**, **DeRamus, M.**, **Wheeler, A. C.**, & **McCormack, L.** (2018). Mobile technology use and skills among individuals with fragile X syndrome: implications for healthcare decision making. *Journal of Intellectual Disability Research*, **62**(10), 821–832. https://doi.org/10.1111/jir.12537

R Core Team (2018). *R: A Language and Environment for Statistical Computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Ruano, L., **Sousa, A.**, **Severo, M.**, **Alves, I.**, **Colunas, M.**, **Barreto, R.**, **Mateus, C.**, **Moreira, S.**, **Conde, E.**, **Bento, V.**, **Lunet, N.**, **Pais, J.**, & **Tedim Cruz, V.** (2016). Development of a self-administered web-based test for longitudinal cognitive assessment. *Scientific Reports*, **6**(1), 1–10. doi:10.1038/srep19114

Rust, K., **Malu, M.**, **Anthony, L.**, & **Findlater, L.** (2014, June). Understanding child defined gestures and children's mental models for touch screen tabletop interaction. In Proceedings of the *2014 conference on Interaction design and children* (pp.201–204). ACM.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, **78**(1), 153–189.

Spaulding, T. J., **Plante, E.**, & **Farinella, K. A.** (2006). Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools* **37**, 61–72.

Stock, S. E., **Davies, D. K.**, & **Wehmeyer, M. L.** (2004). Internet-based multimedia tests and surveys for individuals with Intellectual Disabilities. *Journal of Special Education Technology*, **19**(4), 43–47. https://doi.org/10.1177/016264340401900405

Tomblin, J. B., **Records, N. L.**, & **Zhang, X.** (1996). A system for the diagnosis of specific language impairment in kindergarten children. *Journal of Speech Language and. Hearing Research*, **39**, 1284–1294. doi: 10.1044/jshr.3906.1284

Wood, E., **Petkovski, M.**, **Pasquale, D. D.**, **Gottardo, A.**, **Evans, M. A.**, & **Savage, R. S.** (2016). Parent scaffolding of young children when engaged with mobile technology. *Frontiers in Psychology*, **7**, 690. https://doi.org/10.3389/fpsyg.2016.00690

Yue, Z., **Barker, J.**, **Christensen, H.**, **McKean, C.**, **Ashton, E.**, **Wren, Y.**, **Gadgil, S.**, & **Bright, R.** (2021). Parental spoken scaffolding and narrative skills in crowd-sourced storytelling samples of young children. In 22nd *Annual Conference of the International Speech Communication Association, Interspeech* (pp. 236–240). (Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH; Vol. **1**). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2021-1297

Cite this article: Frizelle, P., Buckley, A., Biancone, T., Ceroni, A., Dahly, D., Fletcher, P., Bishop, D.V.M., & **Mckean, C.** (2023). How reliable is assessment of children's sentence comprehension using a self-directed app? A comparison of supported versus independent use. *Journal of Child Language* 1–29, https://doi.org/10.1017/S0305000923000545