Asking the Teachers: A Delphi Study on the Selection of Skills and Behaviours for an Assessment of Barriers to Learning for Pupils on the Autism Spectrum with Intellectual Disabilities

Accessible Summary

- Barriers make it hard for pupils to learn. We wanted to assess barriers to learning. We made an assessment called the Assessment of Barriers to Learning in Education – Autism (ABLE-Autism)

- We asked teachers to tell us about the barriers to learning. They told us about lots of barriers. We wanted to include the most important barriers.

- We used the Delphi method with teachers. The Delphi method helps people to agree what is important.

- Teachers told us about the most important barriers to learning. We included the most important barriers to learning in our assessment.

Abstract

Background. This study outlines how a modified Delphi procedure was used with special educational needs teachers to select skills and behaviours for inclusion in the Assessment of Barriers to Learning in Education – Autism (ABLE-Autism). The ABLE-Autism is a new teacher assessment to show progress in barriers to learning for pupils on the autism spectrum with coexisting intellectual disabilities in special schools. The research aim was to select items for inclusion in the assessment based on teacher ratings of relevance, comprehensibility, and comprehensiveness.

Method. Following a review of the literature and teacher focus groups, a list of 86 items was developed and a modified two-round Delphi exercise was conducted with special needs teachers. Items were selected for inclusion if at least 80% of teachers agreed that the item was (a) able to be understood, (b) important to assess, and (c) the median score associated with whether the item was able to be understood and important to assess was 1 (the highest score possible).

Results. In the first round, 56 items met the criteria for inclusion and were retained. Thirty items were amended and, after three items were amalgamated with other items, 27 amended items were included in the second round. After the second round, 14 additional items met the threshold to be included in the final assessment.

Conclusions. After both rounds, 70 items were endorsed by teachers and included in the ABLE-Autism. The input of special needs teachers provided initial face and content validity for the new outcome measure.

Keywords: *Autism, Education, Intellectual disability, Learning (intellectual) disabilities, Special educational needs/disabilities*

Background

Special educational needs (SEN) teachers who work with children on the autism spectrum with coexisting intellectual disabilities in special schools often tailor their teaching, curriculum and environment to the needs of their pupils. The focus of many SEN teachers who work with autistic pupils with intellectual disabilities is to teach the pupils skills and support them to develop their abilities in areas which may improve their quality of life and perhaps lead towards access to mainstream schools and later vocation and employment (Jordan, 2005). This may include directed teaching of skills such as communication, social skills, pre-learning skills and life skills as well as supporting pupils to find strategies to self-manage emotions and anxieties. As with all teaching, outcome and progress assessments are necessary to support SEN teachers to identify individual pupils' abilities, show areas of strength and difficulty, define targets and measure progress.

When considering the challenges faced by pupils on the autism spectrum with intellectual disabilities, assessments within special schools are more likely to be proxy rather than self-report due to coexisting intellectual disabilities and associated complex needs, including communication difficulties. When new assessment tools are being developed, it is important that stakeholders are provided with an opportunity to comment on the assessment content to ensure face and content validity (Boateng, Neilands, Frongillo, Melgar-Quiñonez, and Young, 2018). Content validity, which ensures that the content of the area to be assessed is sufficiently covered, is arguably one of the most important measurement properties of an assessment (Prinsen et al., 2018; Mokkink et al., 2018; Terwee et al., 2018). Descriptions of skills or behaviours included in assessments (i.e. the assessment items) must be relevant, comprehensible and comprehensive in respect of both the domain or construct to be measured and the population in question (Prinsen et al., 2018; Mokkink et al., 2018; Terwee et al., 2018). Rigorous development procedures are important for measurement tools across

contexts and settings, including within education. Content validity is usually evaluated by experts in the relevant areas. In educational assessment, it would be appropriate to consider teachers as experts in the content they are teaching and, as the administrators, interpreters and scorers of the assessment, they should be consulted on the content of assessments which they use.

A recently completed systematic review (Howell, Bradshaw and Langdon, 2020) found that assessments which can be used by teachers in special schools are often developed without input from the teachers who use them. The idea that assessments sometimes "end up measuring things that they can measure more easily" rather than the skills prioritised by teachers and parents affects the content validity of assessments and it is important that appropriate measures are taken at the development and evaluation stages to ensure that assessments are useful and effective (Resnick, 2017, p. 151). The systematic review found 42 studies reporting on 26 assessment tools which assessed outcomes in adaptive behaviour, challenging behaviour and autism-related behaviour for children with intellectual disabilities (Howell, Bradshaw and Langdon, 2020). The review considered aspects of assessments which made them appropriate for use by teachers in special schools in the UK to assess outcomes for children on the autism spectrum with intellectual disabilities. Relevant factors which were considered when evaluating the appropriateness of an assessment included the purpose and population the assessment was developed for, whether teachers had been involved in the development, whether it had been evaluated in an appropriate setting with a relevant population and teacher respondents and the reported measurement properties (Howell, Bradshaw and Langdon, 2020). None of the 26 assessments were found to have met all of these criteria. Eight were developed specifically to measure outcomes for pupils on the autism spectrum, however only three of these were developed for use in schools or evaluated using teacher respondents (Autism Treatment Evaluation Checklist; Pervasive Development

Disorder – Behavior Inventory; Teacher Autism Progress Scale). Of these, only the Autism

Treatment Evaluation Checklist (ATEC) was evaluated in the UK. The responsiveness

evaluation of the ATEC suggested that the assessment showed change over a longer period of

time but indicated that the assessment may not show change over the shorter time period that

teachers may wish to assess progress (i.e. termly or half-termly). Other assessments were

found to be potentially useful but required further evaluation of their measurement properties

with autistic pupils with coexisting intellectual disabilities in the UK. The review concluded

that, depending upon the required purposes and content, there is a need for robust, autism-

specific assessments for use specifically for autistic pupils in special schools that have been

developed with input from teachers. Recent changes to statutory assessment in the UK have

given special schools freedom to select and use assessments which they consider to be

appropriate for the needs of their pupils in areas outside of core curriculum subjects

(Standards and Testing Agency, 2020a; Standards and Testing Agency, 2020b). Therefore, if

teachers identify assessment needs related to particular skills and behaviours for specific

groups of pupils, there is the opportunity to develop robust assessments for the required

purposes.

In response to the systematic review, this paper outlines part of the initial

development of the Assessment of Barriers to Learning in Education – Autism (ABLE-

Autism). The ABLE-Autism is an assessment for SEN teachers to measure progress in

barriers to learning for pupils on the autism spectrum with intellectual disabilities in special

schools. Following the systematic review to identify assessments appropriate for use in

schools (Howell, Bradshaw and Langdon, 2020), focus groups were conducted with SEN

teachers to identify assessment needs, define the domain in question ('barriers to learning' for

pupils on the autism spectrum with intellectual disabilities in special schools) and to identify

the relevant skills and behaviours (Howell, Bradshaw and Langdon, in press). A list of items was then created based on the results of the literature review and focus groups.

A group of SEN teachers were invited to take part in a modified two-round online Delphi exercise in order to examine the set of proposed assessment items. The aims of this research were to consider: (a) the relevance of potential assessment items, (b) the comprehensibility of those items and (c) the comprehensiveness of the list of items by addressing whether any skills or behaviours were missing from the list. Amendments to items were made where necessary and each item was either accepted for inclusion in the assessment or rejected. The three areas of comprehensibility, comprehensiveness and relevance were selected to ensure sufficient content validity as suggested by the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) standards for measurement properties of outcome assessments (Prinsen et al., 2018; Mokkink et al., 2018; Terwee et al., 2018).

Methods

Epistemological Position

The Delphi exercise is a research method based on Dewey's pragmatism which has a focus on research informing practice (Brady, 2015). Pragmatism has been posited as a theoretical position which bridges the two paradigms of interpretivism and positivism (Johnson and Onwuegbuzie, 2004). Positivism or realism has an objective research focus with a view to find the truth or reality and is associated with quantitative research methods. Constructivist or interpretivist paradigms, in contrast, aim to understand the subjective realities that individuals or groups construct and usually involve more qualitative methods and analysis (Brundrett and Rhodes, 2013). Johnson and Onwuegbuzie (2004) reasoned that proponents for the different paradigms each reject the other, resulting in research methods being restricted by an

attachment to one epistemological position, effectively creating barriers to potentially effective research methods. They looked to the ideas of classical pragmatists Charles Sanders Pierce, William James and John Dewey to conclude that the empirical and practical consequences should be considered when judging and selecting research methods and making decisions about next steps. Their advocacy on the pragmatist perspective focused upon its practical *usefulness* in research and identified it as a "middle position philosophically and methodologically" (Johnson and Onwuegbuzie, 2004, p.17). Pragmatism is appropriate for the current research as there is a key focus throughout that the assessment is useful for teachers in a classroom. A Delphi study, as a pragmatic research method, allows groups of experts to reach a consensus, seeking the 'truth' yet accounting for the context of this 'truth' by acknowledging the experiences and subjectivity of the participants. Alongside this, including a qualitative element in the Delphi exercise responses is in line with a pragmatic position as deeper and richer data not only allows for a greater understanding of reasoning behind responses but also reflects practical inquiry which iteratively leads to further action and avoids "simplistic responses to very complex issues" (Yvonne Feilzer, 2010, p.11).

Research Design

The purpose of the Delphi exercise is for the group of participants to anonymously indicate agreement via questionnaire and then refine their answers in successive rounds, driving towards a consensus on the issues in question (Dalkey and Helmer, 1963). The Delphi exercise is an established method of evaluating content validity of assessments by determining consensus of experts on the items and content (Boateng et al., 2018). The Delphi method was chosen for the current research as the focused and quantitative data which are produced for each assessment item allowed for clear decision making about the rejection or retention of items. It is a useful method for bridging the gap between research and practice and, with a focus of this research on the involvement of teachers in assessment development

to ensure validity and utility, the Delphi method was considered suitable (Stone Fish and Busby, 2005). The Delphi method was also considered convenient for teacher participants as they could contribute anonymously, did not have to be physically present and could complete the questionnaires at a time and place of their choosing (Green, 2014). The original Delphi method has been adapted and modified for use in a range of disciplines and covering a wide variety of research designs and aims (Hasson and Keeney, 2011).

Participants

Purposive sampling was used to recruit teachers to participate in the online Delphi exercise. Participants were recruited by emailing teachers who had taken part in focus groups within a previous stage of the research (Howell, Bradshaw and Langdon, in press) as well as teachers in the researchers' professional networks. Participant expertise is important within a Delphi exercise, and dependent upon the purpose, different stakeholders can be regarded as 'experts'. For example, patients rather than clinicians have been used in health-related Delphi studies and are considered "experts by experience" (Law and Morrison, 2014, p. 1348). The eligibility criteria for the current study were: (a) that the participants were qualified teachers, (b) currently working in a special school and (c) had recent experience of working with pupils on the autism spectrum. As a school assessment is likely to be used by all teachers regardless of experience or expertise, it was felt that criteria for further selection (e.g. for a teacher to have been teaching for a specific amount of time) would potentially affect the results and, therefore, no further criteria were required. Eleven participants took part in round one and nine in round two. There is no general consensus for the number of participants in a Delphi exercise. Minimum suggested numbers are usually around 7-10 (Day and Bobeva, 2005) but Delphi exercises have been conducted with as few as three participants (Skulmoski, Hartman and Krahn, 2007).

Other stakeholders such as teaching assistants, parents/families and autistic individuals themselves were considered as participants to provide valuable input into the content of a new assessment tool. At this stage in the assessment development, it was decided that teachers would be consulted due to their wider training, knowledge and experience in assessment, curriculum, specific interventions and pedagogy. As the professionals who will be regularly using, facilitating and interpreting the assessment, they were considered the starting point for the initial selection of items. Additionally, parents and the pupils themselves are not usually directly consulted on the content of educational assessments used in mainstream, special needs, primary or secondary schools. Assessment content in the context of pupils on the autism spectrum with intellectual disabilities in special schools, however, often relates to future outcomes in adulthood and quality of life. Therefore, it is appropriate to gain the views of other stakeholders and this may be a consideration for further evaluation of the assessment.

Procedure

A list of items was drawn up based upon barriers to learning for pupils on the autism spectrum with intellectual disabilities which were identified by two focus groups (Howell, Bradshaw and Langdon, in press) along with an extensive review of the literature on adaptive behaviour, challenging behaviour and autism related behaviour assessments as described in the introduction (Howell, Bradshaw and Langdon, 2020). The items were grouped into five categories and labelled as Learning Behaviour Barriers, Restricted and Repetitive Behaviour Barriers, Emotion and Behaviour Regulation Barriers, Sensory Barriers and Functional Communication Barriers.

Both rounds of the online questionnaire were distributed via Qualtrics. All procedures were in accordance with the ethical standards and approval of the Tizard Ethics Committee and with

the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent and assent was obtained electronically from all individual participants included in the study.

Modifications

There were a number of differences and modifications to the established Delphi method which were made for the purposes of this research and these modifications included the number of rounds and a change of response in the second round questions. Although a three-round Delphi exercise is considered typical, two rounds or even one round may be sufficient dependent upon the research (Skulmoski et al., 2007). The first round of a Delphi is often qualitative. As extensive qualitative data had been collected in order to devise the item list from the two focus groups in the previous stage (Howell, Bradshaw and Langdon, in press), it was felt that the first round of this Delphi exercise could begin with more of a focus on directly rating the relevance, comprehensibility and comprehensiveness of the items. In addition, a two-round Delphi reduced burden on teacher participants instead of requiring them to rate and comment upon the same items three or more times (Walker, Worth and Van den Brande, 2019). Further, with a dropout rate of 18% in round two and with high number of items suggestive of a significantly lower response rate (Gargon, Crew, Burnside and Williamson, 2019), we did not want to risk compromising the quality of the data and subsequent results through a particularly low response rate in subsequent rounds. Delphi exercises may continue until a consensus is reached and, with over 80% of the original items reaching the high threshold for consensus in order to be retained in their original or amended format after two rounds, it was determined that a third round was unlikely to provide any significant change in the overall consensus for remaining items. The two-round Delphi, therefore, was sufficient in order to answer the research questions and establish consensus on the items to be included in the ABLE-Autism.

A second modification to the Delphi procedure was the decision to amend the response options to 'yes' or 'no' for the second round. A traditional Delphi exercise provides participants in subsequent rounds with the same questions in the same format and asks them to reconsider previous responses in light of feedback from preceding rounds. However, the response options for the second round questions were amended in this modified Delphi for the following reason. The spread of scores were low in round one with responses on only eight of the 86 items having a standard deviation (SD) $\geq$ 1 for comprehensibility and 11 items with a SD $\geq$ 1 for relevance, with the highest SD of any item being 1.3. These first round results included more positive and less spread responses than originally expected. In order to differentiate between items which had a strong consensus on inclusion and a moderate consensus, a high threshold was decided upon when defining consensus. The first round responses showed a strong consensus for the inclusion of certain items which were retained after round one and therefore not included in round two. However, the results did not show a similar obvious consensus for exclusion (e.g. all but one item which did not reach the high threshold set for inclusion still received over 50% of the responses suggesting they were, to some degree, relevant and comprehensive). As the decision to include or exclude items was ultimately dichotomous and the nuance of a scale did not provide strong evidence of either inclusion or exclusion, the decision for participants to clearly indicate inclusion or exclusion using a dichotomous response was taken. The decision to change the response options in round two was also influenced by the information on teacher workload above, as participant burden and fatigue were considered if participants were asked to complete the same questions on the same scale a second time.

Round One

Following a pilot test, round one of the modified Delphi exercise contained 86 items and asked participants to rate each item on comprehensibility (how well they understood the

item) and relevance (how important they thought the item was to assess) specifically in relation to pupils on the autism spectrum with intellectual disabilities. Items were rated for each question on a four-point Likert scale ranging from 1 (i.e. very easy to understand or very important) to 4 (i.e. not at all easy to understand or not at all important). As the end result is the dichotomous inclusion or omission of an item in the final assessment, a mid-point or neutral value was not used in order to avoid central tendency bias, social desirability bias and the distortion of data (Nadler, Weston and Voyles, 2015). At the end of each section an open-ended question on comprehensiveness asked if any skills or behaviours in this area were missed or should be added to each of the five categories. A final open-ended question on comprehensiveness at the end of the questionnaire asked for any further comments about the items or assessment as a whole. An optional comment box was also provided for each item when participants could add comments if they wished. As a result of the earlier focus group discussion, items were worded positively (i.e. describing what the pupil can do rather than what they cannot) and potential personalisation of the assessment was outlined by having the items refer to a fictional child, 'Charlie'. The first round of the questionnaire was left open over a period of six weeks as its distribution coincided with the school summer holidays and teachers who may not be available as they would be in term time were given an adequate amount of time to respond. Participants were twice informed by email of the extended time allowed to complete the questionnaire. Eleven responses were received in total. In the first round, 4 questions had one missing response each.

Round Two

The participants who had completed and provided consent in the first round were given access to the second round questionnaire. The consent form and participant information sheet were available if participants wanted the opportunity to review them. Participants were provided with feedback on the first round responses and informed that the items included in

the second round questionnaire were the items which less than 80% of teachers rated as either able to be understood or important to assess. The second round questionnaire contained 27 items which had been amended after feedback from the first round. The second round featured 'yes' or 'no' responses to the two questions on comprehensibility and relevance. As a number of items had already been retained and were therefore not included in the second round questionnaire, the question on comprehensiveness was removed so that participants did not request skills and behaviours which had already been retained with the intention of being included in the assessment. Again, an optional comment box was provided for each item where participants could comment as well as a final, open-ended question at the end of the questionnaire where participants could opt to make any further comments. The second round questionnaire was open for two weeks and participants were reminded by email when the questionnaire was approaching the date of closure. Nine responses were received with a second round response rate of 82%.

Data Analysis

The criteria used in the analysis of data collected through the Delphi method may vary. Hsu and Sandford (2007) recognised that for a four-point Likert scale, at least 70% of respondents needed to rate on one side of the scale with a median of at least 3.25 (or 1.75 or lower for reverse scoring). They identified that a median is an appropriate score to consider in a Delphi exercise as it shows converging opinion. As the current modified Delphi exercise considered comprehensibility, a high threshold for consensus was considered necessary and it was decided that an 80% consensus was deemed appropriate for item retention in this study (McConachie et al., 2018). The criteria used for consensus in round one was that: (a) 80% of participants rated the item as 1 or 2 for comprehensibility (very easy or quite easy to understand), (b) 80% of participants rated the item 1 or 2 for relevance (very important or quite important to assess) and (c) the median score for both questions was 1 (the top

response). If the item met all three criteria then consensus was determined and the item was retained in its current form. Any items not reaching these criteria were eligible for amendment and then returned to the participants in round two.

A similar threshold for consensus was used in round two; that 80% or more of participants identified that (a) 'yes' they understood the item, and (b) 'yes' it is important to assess progress in the skill or behaviour.

Results

A summary of the findings from both Round 1 and 2 is outlined in Figure 1.

Figure 1.

Flow Chart of Results From Round 1 and 2 of the Delphi Exercise.

[INSERT FIGURE 1 HERE]

First Round Results

Retained items

Eighty-six items were presented in round one, and a consensus was reached on 56 items as per the criteria specified above; these items were retained without any changes. For two of these items, 100% of participants answered 1 or 2 to both comprehensibility and relevance.

Items for Amendment

In total, 30 items were considered for inclusion in round two. Twenty-six items had less than 80% of participants rate either comprehensibility or relevance as 1 or 2, therefore not meeting criteria (a) or (b) for inclusion after round one. Three additional items had over 80% of participants rate 1 or 2 for both questions but, with a median score of 2, did not meet

criterion (c). All of these 29 items were set aside for amendment before being taken forward to the next round. One further item was rated as able to be understood by 100% of participants (1 or 2 for comprehensibility) but only 75% of participants rated 1 or 2 for relevance meaning that it did not reach criterion (b). As it was considered comprehensible by all participants, this item was not amended but put back into the round two questionnaire to determine whether a consensus could be reached.

Additional Comments

Eight of the eleven participants made at least one comment in the optional comment boxes in round one. Three teachers suggested additional areas to include. Skills such as "accepting a change or when things don't go according to plan" were identified by participants in the comment box at the end of early sections, however these skills were covered in later sections. Some of the items which were taken forward to the second round were amended to clarify or include suggestions from teachers. For example, the suggestion of including working with peers in different roles was incorporated when the item *Charlie will sit for the duration of a group activity* was amended to *Charlie will appropriately engage in a group activity with peers*.

Further comments on specific items often addressed the scope of the item and asked for more information such as prompt levels. As an aspect of personalisation is intended to be incorporated in the teacher responses in the final version of the assessment, items were not amended to be more specific at this stage and this will be considered further in the discussion section below. The intentions behind some comments were not entirely clear in terms of the survey questions, e.g. one comment stated that "this is very difficult to teach non vocal children". It was unclear how to act upon these comments but, where possible, they were taken into consideration when amending the items for round two.

Second Round Results

Items which did not meet the criteria for retention after round one had their wording amended for inclusion in round two. It was determined that three of the items could be amalgamated with other items after rewording. For example, the item *Charlie will show awareness of different expectations in different rooms/settings* and the item *Charlie will show awareness of different behaviours being appropriate/inappropriate depending on the setting* were joined to form the amended item *Charlie will adjust their behaviour depending upon the expectations of the environment or setting*. In total, 27 items were presented in the round two questionnaire.

Retained Items

Responses from the second round indicated that 25 items were understood by 80% or more of participants. Fourteen of these items were considered to be important to assess by 80% or more of participants and, therefore, were retained in their amended version for the final version of the assessment. This led to a total of 70 items to be included within the ABLE-Autism.

Additional Comments

Five participants made optional comments which covered varying interpretations, positive feedback on the items and stipulations on items (e.g. "yes if meaningful"). As in the first round, a number of comments related to the specific level of support or interpretation around language used within the item. Again, this is intended to be addressed through aspects of personalisation in the final version of the assessment but is an issue, as recognised in Howell, Bradshaw and Langdon (in press), that will need further consideration.

Discussion

The modified Delphi exercise identified a total of 70 items which are to be included in the ABLE-Autism. Teachers rated the original list of items, devised from the literature and teacher focus groups, for comprehensibility, relevance and comprehensiveness as suggested by the COSMIN standards for measurement properties of outcome assessments (Prinsen et al., 2018; Mokkink et al., 2018; Terwee et al., 2018). Showing a consensus in these areas provides preliminary evidence for face and content validity of an assessment (Boateng et al., 2018). Sixty-five percent of items were retained without change in the first round of the modified Delphi exercise and 52% of items presented in the second round were retained. The final list of items selected for inclusion from this modified Delphi study consisted of 14 Learning Behaviour Barrier items, 12 Restricted and Repetitive Behaviour Barrier items, 16 Emotion and Behaviour Regulation Barrier items, 10 Sensory Barrier items and 18 Functional Communication Barrier items.

The percentage of items retained was reasonably consistent across categories. 67% (14 of 21) of the Learning Behaviour Barrier items were included in the final assessment after both rounds of the modified Delphi exercise, with 77% of items (10 of 13) included in the Sensory Barriers, 84% of items (16 of 19) included in the Emotion and Behaviour Regulation Barriers category and 86% of items (12 of 14) included in the Restricted and Repetitive Behaviour Barriers category. Only one item from the set of 19 Functional Communication Barrier items did not reach the threshold for inclusion in the first round or in its amended form in the second round (95%). The fact that no single category contained a high percentage of rejected items suggested that the five categories reflect areas which teachers believe are important domains when assessing pupils on the autism spectrum with intellectual disabilities in special schools. The Functional Communication Barrier category contained the largest number of retained items which suggested that SEN teachers rate a variety of functional

communication skills as important to assess when considering and reporting pupil progress in schools.

Thirteen items were not considered relevant for assessment by a sufficient percentage of teachers. Items relating to showing pride or enjoyment in completing an activity and tolerating sensory input were rated among the least important skills or behaviours to assess in spite of these areas being brought up in teachers' focus group discussions as important areas of progress. It may be that, as discussed in Howell, Bradshaw and Langdon (in press), the practicalities of translating these skills or behaviours into items to assess pupils means that the skills are difficult to capture. Similarly, the comments by teachers on possible interpretations of words included in the assessment also reflects teachers' concerns around ensuring that descriptions are precise to avoid variations in the understanding of assessment items. The balance between assessments being specific enough to ensure reliability yet flexible to allow for the ways different pupils may display a particular behaviour or skill is difficult to strike when considering the wording and content of assessment items.

Six of the items retained in round one received one score of 4 for comprehensibility (not at all easy to understand). Although these items were included as they exceeded the 80% threshold for both questions with a median of 1, they will be specifically marked to ensure they are checked further for comprehensibility in later evaluation studies.

The main practical implication of this Delphi study is the development of a new teacher assessment. The assessment has a number of aims and uses from a school and teacher perspective. Firstly, the assessment will allow SEN teachers to identify pupils' barriers to learning or to formally and systematically record existing barriers which have already been identified by the teacher. It may help teachers ascertain specific skills, areas or behaviours which may benefit from individualised and targeted support, teaching, intervention and

environmental adaption. It will enable teachers to measure and record change in these areas. This will support teachers to evidence when a pupil is making progress in important and relevant skills and behaviours, even if progress is not being made in curriculum learning.

Teachers have been and will continue to be involved in the development and evaluation of the ABLE-Autism. Statutory assessments used in special schools are developed to assess pupils with a wide range of intellectual and developmental disabilities and are prescribed by government. The ABLE-Autism is an assessment that directly addresses skills and behaviours that SEN teachers have identified as priorities for teaching specifically for pupils on the autism spectrum with intellectual disabilities and also incorporates areas which SEN teachers have recognised as difficult to assess. The recent systematic review identified few existing assessments which were developed with input from teachers and evaluated in a UK special school context with these specific pupils (Howell, Bradshaw and Langdon, 2020). Of the assessments that may be appropriate for these purposes, such as the Autism Treatment Evaluation Checklist, responsiveness evaluations have suggested that the assessments may show progress over a longer time period (e.g. a year), whereas teachers have identified that sometimes they may need to record small steps of progress more regularly, particularly in areas that are barriers to a pupil's learning (Howell, Bradshaw and Langdon, in press). Allowing teachers to comment on potential items, suggest amendments and further inclusions and make decisions on the relevance of items, makes it likely that the items will show the areas of progress in the ways that teachers will find useful.

This research also shows that teachers can give valuable input into the assessment development process through use of modified Delphi procedures and this will support the content validity of the assessment in future evaluations. It is likely that an assessment will be more useful and valuable if it was developed for the needs of the specific group who will be assessed, if it addresses the relevant areas and was developed and evaluated in consultation

with those who will use it. Further evaluation of the ABLE-Autism may provide further evidence that teacher input supports the development of valid, reliable and useful assessments for use by teachers in special schools.

Limitations

This modified Delphi exercise enabled the selection of a number of different skills and behaviours to be included in the ABLE-Autism. There are, however, some limitations of the current research.

The sample was small and, although Delphi exercises can be conducted with fewer participants (Day and Bobeva, 2005), a larger population of teachers would have been preferable. A dropout rate of 18% from round one to round two is likely to have had a greater effect on the results when the initial sample is small, although second round responses appeared broadly in line with the first round responses and the results between rounds did not seem to alter in an unexpected way. The small number of participants choosing to participate in the modified Delphi exercise may be due to a number of factors including teacher workload, the initial length of the questionnaire and the fact that the availability of the questionnaire coincided with part of the school holidays. It is important to note that, although a Delphi exercise is considered a way for expert opinion to reach consensus, the responses from the group of participants may not be generalisable to the entire population being considered (Skulmoski et al., 2007). It is necessary to consider that the results may not entirely reflect the views of all teachers and, therefore, further evaluation of the items in the finalised versions of the assessment is recommended with a larger sample. Similarly, this study considered teachers to be 'experts by experience'. The use of participants with specified experience relating to years teaching, level of responsibilities or teaching achievements was considered to conflict with the fact that the assessment is intended to be

useful to everyday classroom teachers, regardless of experience. Therefore, participants'
'expertise' in the relevant area was determined by the fact that they were qualified teachers
currently working in a special school and had recent experience of working with autistic
pupils. This is also a limitation as, although this was not believed to be directly relevant for
participant selection in the current study, it may be that newly qualified teachers have a
different interpretation of important skills to assess to more experienced teachers. As
suggested by Hasson and Keeney (2011), Delphi studies should not be taken as "indisputable
fact" and instead be accepted as the opinions of a particular group at a particular time (p.
1701).

Another limitation of the study is that, although other stakeholders were considered
for participation, they were not consulted at this stage. Consulting families, other
professionals and the pupils themselves where possible would be a useful way to triangulate
the findings in the early stages of the development of this assessment. It would also ensure
that the assessment content adequately covers important quality of life and family quality of
life outcomes as well as targeting areas leading to good outcomes in adulthood and future life
opportunities. Evaluation and validation of an assessment is an ongoing process and,
therefore, consulting and including these stakeholders in assessment evaluation is
recommended in future.

A further limitation of a Delphi exercise, particularly a mini-Delphi consisting of only
two rounds, is the fact that misunderstandings may arise (Pan, Vega, Vella, Archer and
Parlett, 1996). Participants were able to contact the researcher by email or phone at any time
if they had questions, though none did so. Certain comments were not able to be acted upon
as they were not clear and the responses could not be followed up and clarified due to the
anonymity of participants, although these were often from only one participant and ratings of
the items supporting their opinion were taken account of within the data analysis.

Conclusion and Recommendations

This paper has outlined the use of a modified Delphi method to identify a total of 70 items endorsed by SEN teachers which will be included in the ABLE-Autism. These 70 items will initially be divided into five categories or subscales. The use of the modified Delphi method to select the appropriate items to assess barriers to learning for autistic pupils with intellectual disabilities in special schools provides initial face and content validity for the assessment. It is recommended that an evaluation of these items and the final version of the assessment as a whole is conducted using a relevant population in a school setting with teacher respondents in order to ascertain the measurement properties of the ABLE-Autism. It is also recommended that, in future, other stakeholders such as teaching assistants, parents, families and pupils themselves are included in further evaluation of the assessment content and uses. As well as evaluation of the reliability and validity of the ABLE-Autism, it is necessary to continue to check content validity with teachers once the assessment has been used to ensure that the assessment is useful in practice as well as in theory.

Author Note

A copy of the assessment can be provided upon contacting the authors.

Conflict of Interest

The authors declare that they have no conflict of interest.

Data Availability Statement

Data will be deposited in the Kent Academic Repository (https://kar.kent.ac.uk/) which can be accessed upon request. The ABLE-Autism can be made available on request.

References

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in Public Health*, *6*. doi:10.3389/fpubh.2018.00149

Brady, S. R. (2015). Utilizing and adapting the Delphi method for use in qualitative research. *International Journal of Qualitative Methods*, *14*(5). doi: 10.1177/1609406915621381

Brundrett, M., & Rhodes, C. (2013). Theories of educational research. *Researching Educational Leadership and Management, Sage Publications Ltd, London*, 12-23.

Dalkey, N., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science*, *9*(3), 458-467.

Day, J., & Bobeva, M. (2005). A generic toolkit for the successful management of Delphi studies. *The Electronic Journal of Business Research Methodology*, *3*(2), 103-116.

Gargon, E., Crew, R., Burnside, G., & Williamson, P. R. (2019). Higher number of items associated with significantly lower response rates in COS Delphi surveys. *Journal of clinical epidemiology*, *108*, 110-120.

Green, R. A. (2014). The Delphi technique in educational research. *Sage Open*, *4*(2),1-8. doi:10.1177/2158244014529773

Hasson, F., & Keeney, S. (2011). Enhancing rigour in the Delphi technique research. *Technological Forecasting and Social Change*, *78*(9), 1695-1704. doi:10.1016/j.techfore.2011.04.005

Howell, M., Bradshaw, J., & Langdon, P. E. (2020). A Systematic Review of Behaviour-Related Outcome Assessments for Children on the Autism Spectrum with

Intellectual Disabilities in Education Settings. *Review Journal of Autism and Developmental Disorders*. https://doi.org/10.1007/s40489-020-00205-y

Hsu, C. C., & Sandford, B. A. (2007). The Delphi technique: making sense of consensus. *Practical Assessment, Research & Evaluation.* 12(10), 1-8.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, *33*(7), 14-26.

Jordan, R. (2005). *Autism with Severe Learning Difficulties*. London: Souvenir Press (Educational and Academic) Ltd.

Law, H., & Morrison, A. P. (2014). Recovery in psychosis: a Delphi study with experts by experience. *Schizophrenia Bulletin*. 40(6), 1347-1355. doi:10.1093/schbul/sbu047

McConachie, H., Mason, D., Parr, J. R., Garland, D., Wilson, C., & Rodgers, J. (2018). Enhancing the validity of a quality of life measure for autistic people. *Journal of Autism and Developmental Disorders*, *48*(5), 1596-1611. https://doi.org/10.1007/s10803-017-3402-z

Mokkink, L. B., De Vet, H. C., Prinsen, C. A., Patrick, D. L., Alonso, J., Bouter, L. M. & Terwee, C. B. (2018). "COSMIN risk of Bias checklist for systematic reviews of patient-reported outcome measures." *Quality of Life Research*, *27*(5), 1171-1179.

Nadler, J. T., Weston, R., & Voyles, E. C. (2015). Stuck in the middle: the use and interpretation of mid-points in items on questionnaires. *The Journal of General Psychology*, *142*(2), 71-89. doi: 10.1080/00221309.2014.994590

Pan, S. Q., Vega, M., Vella, A. J., Archer, B. H., & Parlett, G. R. (1996). A mini-Delphi approach: An improvement on single round techniques. *Progress in Tourism and Hospitality Research*, *2*(1), 27-39.

Prinsen, C. A., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., De Vet, H. C. & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, *27*(5), 1147-1157.

Resnick, M. (2017) *Lifelong Kindergarten: Cultivating Creativity through Projects, Passion, Peers and Play.* Massachusetts: MIT Press.

Skulmoski, G. J., Hartman, F. T., & Krahn, J. (2007). The Delphi method for graduate research. *Journal of Information Technology Education: Research*, *6*(1), 1-21.

Standards and Testing Agency. (2020a). *Pre-key stage 1: pupils working below the national curriculum assessment standard for the 2020/21 academic year onwards.* London: Department for Education.

Standards and Testing Agency. (2020b). *Pre-key stage 2: pupils working below the national curriculum assessment standard for the 2020/21 academic year onwards.* London: Department for Education.

Stone Fish, L., & Busby, D. M. (2005). The Delphi method. In Sprenkle, D. H., & Piercy, F. P. (Eds.). *Research methods in family therapy (pp238-253).* London: Guildford Press.

Terwee, C. B., Prinsen, C. A., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J. & Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Quality of Life Research*, *27*(5), 1159-1170.

Walker, M., Worth, J., & Van den Brande, J. (2019). *Teacher Workload Survey 2019.* London: Department for Education.

Yvonne Feilzer, M. (2010). Doing mixed methods research pragmatically: Implications for the rediscovery of pragmatism as a research paradigm. *Journal of mixed methods research*, *4*(1), 6-16.

Data will be deposited in the Kent Academic Repository (https://kar.kent.ac.uk/) which can be accessed upon request (citation to be added).

Figure Legends

Figure 1. Flow Chart of Results From Round 1 and 2 of the Delphi Exercise.