Received: 15 July 2021 | Revised: 23 September 2021 | Accepted: 27 September 2021

DOI: 10.1002/bin.1832

RESEARCH ARTICLE

WILEY

Thematic and textual analysis methods for developing social validity questionnaires in applied behavior analysis

Rachel Anderson[1] | **Sarah Taylor**[1] | **Tessa Taylor**[2] | **Javier Virues-Ortega**[1,3]

[1]The University of Auckland, Auckland, New Zealand

[2]Paediatric Feeding International, Sydney, Australia

[3]Universidad Autónoma de Madrid, Madrid, Spain

Correspondence
Javier Virues-Ortega, Universidad Autónoma de Madrid, Spain.
Email: javier.virues-ortega@uam.es

Funding information
ABA España, Grant/Award Number: CON02739

Abstract

Social validity is often defined as the degree to which an intervention has value to the community that it effects, but it is seldom reported in the literature. Most social validity questionnaires are purposely created by the authors of the study and often lack a description of scale development process. The purpose of this study was to evaluate methods for the development of a Likert-type social validity scale. Caregivers of children who took part in a study on behavioral treatments for pediatric feeding disorders were part of an initial interview to inform scale development. We analyzed interviews using thematic (qualitative) analysis and textual (quantitative) analysis, and used the resulting themes to generate items for two social validity questionnaires. We examined the inter-rater reliability of the questionnaire development process and evaluated the content validity of the questionnaires resulting from each method. Textual analysis had higher inter-rater reliability for producing themes that could be converted to questionnaire items. The textual analysis method produced a questionnaire with content validity equal to that of the thematic analysis method. The study demonstrates the successful use of a quantitative approach

to the development of social validity questionnaires for behavioral interventions.

KEYWORDS
behavior-analytic interventions, pediatric feeding, questionnaire development, social validity

1 | INTRODUCTION

Social validity refers to the extent to which an intervention is valued by those receiving it, their families, and caregivers. Baer et al. (1968) described socially important interventions as those producing not only a change in behavior, but a change of practical importance. They proposed that focusing on socially important behavior change could lead to social acceptance of the practice used to achieve the change. According to Kazdin (1977), a socially valid behavior change is one that is relevant to the individual's natural environment. In addition, Wolf (1978) defined social validity according to three dimensions: (a) socially significant goals; (b) acceptability of procedures; and (c) importance of effects.

Socially significant goals are those that are meaningful to the client and significant others (see also Bailey & Burch, 2018, pp. 67–74). Hawkins (1991) suggested that including goals in social validity assessments may help to predict the level of cooperation in treatment, the likelihood of the client's return to treatment when required, and the likelihood of the treatment being recommended to others. The acceptability of procedures is the extent to which the methods used to achieve the target behavior change are tolerated and valued by those receiving behavioral services. Specifically, the clients and their family must find procedures acceptable. The importance of the effects is often interpreted in terms of direct outcomes of the intervention. It also includes 'side effects' of the intervention. For example, if an intervention with the intention of increasing a child's self-feeding with utensils and oral solid food volume also improved toothbrushing acceptance and verbalizations (i.e., positive side effects).

Social validity has even been described as a distinctive characteristic of applied behavior analysis and it is recommended that social validity is reported alongside intervention effectiveness (Finn & Sladeczek, 2001). However, relatively few intervention studies report social validity. A review of social validity trends in the *Journal of Applied Behavior Analysis (JABA)* and in *Behavior Modification* found that 20% of intervention studies reported social validity (Kennedy, 1992). A similar review by Carr et al. (1999) found that measurement of treatment acceptability among empirical studies published in *JABA* increased after the 1970s from 0 to 10%, reached a peak of 30% in the 1980s, and stabilized at about 16% in the late 1990s. More recently, systematic reviews have continued to emphasize the lack of social validity reporting in a range of intervention studies (Meindl et al., 2019; Wilder et al., 2020).

Kennedy (1992) found that 54% of articles containing a social validity measure focused on the social validity of the effects of the behavior change programs, 25% focused on the social validity of the goals, and 22% focused on the social validity of the procedures. A more recent review by Ferguson et al. (2019) studied social validity publication trends in *JABA* for the period 1999 through 2016. The authors reported a modest increasing trend from around 5% of publications in the late 1990s or early 2000s to around 20% in the mid 2010s. Unlike Kennedy (1992), they found that 85% of the articles reviewed focused on the social validity of the procedures, 60% on the social validity of the effects, and 12% on the social validity of the goals. Thus, researchers tend to assess the social validity of the procedures, while the assessment of the social validity of goals and intervention effects have remained largely unchanged since the early 1990s, according to these analyses.

Different groups of stakeholders can impact on the successful adoption of an intervention model in different ways. Specifically, some groups will identify certain aspects of social validity as being more important than others

(Foster & Mash, 1999). For example, cost may be an important consideration for those who are implementing or subsidizing interventions, whereas side effects may be more important to the recipient's family and caregivers.

Three groups of stakeholders are used to obtain social validity measures: blind assessors, those who interact daily with clients or participants (e.g., parents or teachers), and the recipients of behavioral services themselves. Hanley (2010) recommended obtaining social validity from the recipient of the behavior change program in order to allow the client's preferences to influence the choice of treatment (some would say to *empower* the client). While treatment preference may be possible to assess under most circumstances (see e.g., Hanley et al., 1997, 2005), not all participants can respond appropriately to the questionnaires and interviews often used to assess social validity. In fact, proxy-reporting of social validity is the most common approach to informing social validity, particularly on occasions when the intervention recipients cannot communicate verbally. For example, Kryzak et al. (2013) had parents as social validity informants for a joint attention intervention conducted at home, whereas teachers commented on intervention sessions delivered in the school environment.

1.1 | Social validity questionnaires

For over 30 years, most social validity measures have been obtained via Likert-type questionnaires, which allow for subjective perceptions of the respondent to be analyzed as an interval-level quantitative variable. Many social validity questionnaires are developed specifically (i.e., ad hoc) for a given study, whereby researchers determine which questions are important to be included and rated, given the aims of the study. Authors may develop these questionnaires in relation to those previously used in the literature or may create a novel set of questions. In addition, some standardized assessments have been developed. The Treatment Evaluation Inventory (TEI) was developed to evaluate treatments for problem behaviors (Kazdin, 1980), and was later followed by a shorter version containing nine 5-point Likert-type items (Kelley et al., 1989). The Intervention Rating Profile-20 (IRP-20; Witt & Martens, 1983), the Behavior Intervention Rating Scale (BIRS; Von Brock & Elliott, 1987), and the Abbreviated Acceptability Rating Profile (AARP; Tarnowski & Simonian, 1992) are additional examples. Only one Likert-type scale, the Scale of Treatment Perceptions (STP), has been developed specifically to evaluate treatments that increase desired behaviors (Berger et al., 2016). A potential limitation of using a standardized test for assessing social validity is that it may only be relevant for the specific intervention-population combination for which it was originally developed and validated. For example, a social validity test created with families with high socioeconomical standing may not be relevant if used with less affluent families, even if the intervention and clinical characteristics of clients are comparable. Therefore, ad hoc social validity tools may continue to be valuable in some situations.

Few studies have used open-ended interview questions to measure social validity, which may glean more relevant information on what the treatment means to each individual consumer. For example, Axe and Sainato (2010) asked teachers open-ended questions on the strengths and weaknesses of their intervention to teach multiword phrases to children with autism. They found that teachers frequently voiced concerns about how the intervention would work in a classroom where they could not always provide one-on-one teaching. Some pediatric feeding studies have use open-ended questions anecdotally to assess preference for specific physical guidance procedures (Rubio, McMahon & Volkert, 2020). While open-ended questions may be less biased toward the interventionist's perception of what a socially valid treatment might be, it is challenging to aggregate and scale the information resulting from open-ended questions.

There are no standard methods for reporting the reliability and validity of social validity measures as they are commonly used in applied behavior analysis. According to Schwartz and Baer (1991), social validity measures lack psychometric rigor, and their reliability and validity is unknown. Schwartz and Baer suggest that the internal consistency of social validity questionnaires could be improved by first studying antecedents to the acceptance or rejection of behavioral programs, and then relating this to information obtained from the questionnaires. According to Fuqua and Schwade (1986) most questionnaires used for assessing social validity rely on face validity. The appropriate statistical

tests to establish the stability of data from these questionnaires has, for the most part, never been established. Further research into variables correlated with high scores on social validity questionnaires is needed to establish test-retest reliability for these questionnaires as well as their validity. Fuqua and Schwade also suggest that the predictive validity, that is, the degree to which the behavior being changed is the behavior driving the ratings in the social validity instrument, has not been tested for most questionnaires. Construct validity, established by way of factor analysis, and internal consistency, informed by Cronbach alpha, are the most common psychometric attributes explored during scale development of Likert-type scales (Hinkin, 1998), yet these are seldom reported in social validity questionnaires. Psychometric standards are only available for standardized social validity questionnaires, such as the TEI, and they are often derived from the specific population available when the questionnaire was first developed.

1.2 | Item-generation methods

It is rare for studies using social validity questionnaires to adequately describe the methods used for item generation. For example, the author of the TEI described item generation as derived from "face validity" of aspects that related to client's evaluation of treatment, including relevance to children and the use of punishment (Kazdin, 1980). Further, the authors of the IRP did not describe how items were generated (Witt & Martens, 1983), and authors using purposely created (ad hoc) questionnaires do not describe how they were developed. Overall, current measures of social validity could be strengthened by a more systematic approach to questionnaire development that places more emphasis on the views of the consumers of interventions.

Recent research has proposed alternative methods for item generation, specifically the use of qualitative methods (Leko, 2014). Qualitative analyses are typically concerned with representing how people think and feel in a particular situation delineated by the research question (Thorne, 2000). One qualitative method is a semi-structured interview to allow participants' own spoken words to form descriptive data. During an interview, participants can share detailed individual experiences or perceptions, in their own terms. Qualitative research may therefore increase the potential to identify factors that are most important to participants, and this is key to the broader view of social validity. This descriptive interview data can be synthesized into common themes across participants, that inform social validity questionnaire items (Leko, 2014). This process can be replicated for specific interventions, thereby eliminating the cost of development of standardized tests. Two methods to generate themes from interview data include *thematic analysis* and *textual analysis*. We recognize that these methods are largely unfamiliar to the behavior-analytic community, thus we will attempt to draw on related literature where these methods have been used in areas complementary to behavior analysis.

Thematic analysis involves identifying patterns (themes) from interview transcripts that are important in relation to the research question (Braun & Clarke, 2006). Using the same coding system, themes may be synthesized across interviews from different participants. Overall, thematic analysis is intended to comprehend the perspective of a group of individuals with a similar set of circumstances (Braun & Clarke, 2006; Thorne, 2000). Thematic analysis has been used to identify quality of life indicators for children with autism (Epstein et al., 2017), and the wider benefits to staff after training in positive behavior support (Walsh et al., 2019).

Briefly, thematic analysis procedures involve accurate transcription of interviews, attention to each data item (word), checking and re-checking of themes, and a well-organized story about the topic (see Braun & Clarke, 2006 for a 15-point checklist). Thematic analysis may be an important method for generating themes from open-ended questions that could then be used to generate items for social validity questionnaires. A potential limitation to this approach is that the inter-rater agreement of the process of identifying themes is seldom reported in the literature with some authors arguing that it may not be an adequate metric for qualitative analysis (Armstrong et al., 1997).

Another method for generating themes from interview data is textual analysis. Textual analysis is a collection of systematic quantitative strategies to extract *meaning* from samples of natural language (Landauer et al., 1998). There are a number of computer programs that conduct textual analyses by grouping words into themes that may be seman-

tically relevant. For example, the 3rd Eye theme analysis algorithm by Hunerberg (2019) generates themes by identifying words that tend to co-occur within sentences. Other systems may be able to process the grammatical functions of sample texts, thereby identifying, for example, specific noun-noun and noun-adjective combinations (e.g., treatment-benefit, intervention-positive) that may be found repeatedly in the text analyzed (see e.g., Michel et al., 2010). Yet other language processing systems can explore some basic semantic aspects of a text by grouping together words of similar or related meaning. For example, the words *progress*, *gain*, *develop*, and *achieve* could be grouped into a theme (see e.g., Semantic Knowledge, 2014). These analyses provide the basis for identifying meaningful themes from a sample of natural language (e.g., social validity interview transcript) without requiring potentially arbitrary or biased decisions from a human rater. Therefore, the use of textual analysis tools could help to minimize the sources of bias that have been often recognized in qualitative research methods (see e.g., Galdas, 2017).

1.3 | Social validity in pediatric feeding interventions

In the current study, caregivers of children receiving behavior-analytic treatment for pediatric feeding problems participated in a comprehensive social validity analysis. The few studies in pediatric feeding that have addressed social validity typically used Likert scale questionnaires, including standardized questionnaires (e.g., Hoch, et al., 1994; Sharp et al., 2014, 2016; Taylor et al., 2020). For example, Bui et al. (2014) used the BIRS to evaluate the social validity of nonremoval of the spoon in combination with verbal reinforcement in a child with autism. Wood et al. (2009) used the IRP-15 as a measure of social validity for their intervention with a child food selectivity. Woods and Borrero (2019) used an adapted version of the IRP-15 in an examination of extinction bursts in pediatric feeding treatment. Pediatric feeding specialists have also developed social validity questionnaires for various specific treatment procedures and purposes (Ahearn et al., 1996; Borrero et al., 2013; Kozlowski et al., 2016; Rubio, Volkert, et al., 2020; Taylor, 2020; Ulloa et al., 2019).

The aim of the current study was to evaluate two methods for the development of social validity questionnaires: thematic analysis (i.e., a qualitative approach) and textual analysis (i.e., a quantitative approach). The study aimed to address some of the current limitations in the literature by: (a) including assessment of goals, procedures, and outcomes in the social validity measure *à la* M. Wolf; (b) identifying the appropriate informant to use for social validity assessment; and (c) reporting the process to generate Likert scale items. For the purpose of pilot-testing a specific population, these methods involved open-ended interviews with consumers of behavioral treatments for feeding disorders, to obtain an overall picture of the consumer's thoughts on the treatment. These interviews were then analyzed using (a) thematic analysis, and (b) textual analysis, to uncover themes and synthesize open-ended interview data. The resulting themes were then used as the basis of generating items for a Likert scale questionnaire to evaluate the social validity of behavioral treatments for transitioning children from tube to oral feeding.

2 | METHOD

2.1 | Participants and setting

Participants were recruited from an on-going study evaluating home-based behavioral treatments to transition children from tube to oral feeding (Taylor et al., 2019). The parents of the children receiving the intervention in the original study served as participants in the current analysis. All nine families in the consecutive case series were invited to be part of the current study. Seven families agreed to participate and two declined. Four mothers, one grandmother, and two mother-father dyads participated in the interview. The interview was conducted while the behavioral intervention was on-going for two families, within the week of ceasing treatment for one family, and between 5 and 10 months after the child had completed treatment for the remaining four families. Six participants also completed

the social validity questionnaires resulting from the thematic and textual analyses following the semi-structured interview. The questionnaire required about 15 minutes to complete. A postgraduate student completing a Master's degree in behavior analysis conducted all interviews and analyses. The student was not involved in the behavioral intervention, and families were also informed that their information would be anonymized and not accessed by the board-certified behavior analyst (BCBA) providing the behavioral intervention.

Ethical approval for this study was granted by the local Health and Disability Ethics Committee (New Zealand). Initial contact with potential participants for the current study was made by the researcher of the original study. Each participant was provided with a participant information sheet and consent form which signed before proceeding with the interview.

2.2 | Semi-structured interview

A semi-structured interview was conducted with each participant individually, either in person or by phone call. Semi-structured interviews were selected for investigating social validity because it was possible to explore the participants' perspectives of the treatment with enough flexibility to obtain sufficient detail, without neglecting any unexpected findings.

Interview schedules were developed with reference to the social validity literature (Supplementary Online Information, Appendix A). Namely, we included questions on the goals, procedures, and outcomes of treatment (Wolf, 1978). This included a question on the unpredicted effects of treatment, "Were there any extra effects that the treatment had on your child? These could be beneficial or unhelpful extra effects." Questions were kept as open as possible to allow participants to provide any information that they thought relevant. In addition, probes were included to encourage participants to provide as much detail as possible during the interview. For example, if in response to the question, "How did you feel about the relevance of the treatment goals to your child's situation?" the participant replied "Yes, they were relevant", probe questions would follow to encourage further detail (e.g., "Were they applicable?" "Why/why not?").

All interviews were audio-recorded and transcribed. Each interview was completed within one hour or less. Identifying information was removed and names replaced with pseudonyms during the transcription process. The interviewer provided minimal comment but encouraged the participant to continue speaking and expanding their responses by nodding, expressing agreement, or asking for clarification in the form of probe questions where required.

2.3 | Thematic analysis

We used thematic analysis to produce themes from the participant's answers to each question in the semi-structured interview, according to the guidelines by Braun and Clarke (2006). We used the interview audio recordings to transcribe the answers to all questions. We then created separate documents each containing all the answers for a given interview question. We used codes to signify important pieces of information within the narrative record. For example, "I was thinking that maybe after a year, that maybe Nate start eating [self-feeding]. It wasn't that easy" was coded as "progress was slow." The codes were then grouped into potential themes that captured a central idea, summarized in a single sentence. For example, the code "progress was slow" as well as similar codes (e.g., "it all takes time") were combined to form a theme covering *how parents felt about the time frame and time commitment of the intervention*. Multiple themes were able to be generated from responses to each question, up to a maximum of six themes. Potential themes were reviewed in relation to the dataset, revised, and collated further where necessary in order to form a small group of cohesive themes. Table 1 presents a detailed technological description of the thematic analysis process (additional details are available in the Supplementary Online File, Appendix B).

TABLE 1 Thematic analysis steps

Step	Description
1. Review the data	Read the interview question and responses through to get a general idea of what it is about. Re-read the interview question and responses, then note down any initial impressions of central or recurring topics.
2. Generate initial codes	Carefully go through the interview questions and break down the responses. Interviewer statements are not coded. For any piece of potentially interesting information, create a 'code' or small description in just a few words. Use microsoft Word to highlight small portions of the text such as part of a sentence, include the code within a comment. Pay careful and equal attention to all parts of the text and apply this coding process to a high level of detail.
3. Searching for themes	Create a list of all the codes created. Use this list to collate the codes into groups that could potentially become the themes. Write each code onto a post-it flag or paper then arrange these into groups, on a wall, or desk. Between two and six themes should be generated for each question.
4. Reviewing themes	Check that the themes correspond with the codes and the portions of the text that these codes were created from. Re-read the interview to check that the themes make sense overall. This can help to clarify whether a subgroup should be added. Alternatively, discard small miscellaneous groups that do not form a significant or coherent theme.
5. Defining and naming themes	Create a clear thematic statement in one sentence for each of the themes you extract. It is important that themes are not just repeating the questions or extracts of the interview, but describe what is interesting about the idea and why. Remember that later, these themes could be expanded on and discussed in detail. For example, a final theme could be "expectations—families were cautious about having expectations, and generally did not know what sort of outcome to expect."

2.4 | Textual analysis

Textual analysis was the second method used to extract themes from the interview transcriptions. We used the free software package Tropes® (Semantic Knowledge, 2014) in order to conduct semantic and textual analyses. The Tropes® package was selected because it was appropriate for the analysis of natural language and easy to use. As described for the thematic analysis, we used the interview transcriptions for a given question and imported it onto the software for analysis. The software groups words with related (e.g., battery, power, electricity) or similar meanings (e.g., durable, permanent, solid), words that bear similar moods (e.g., deficiency, failure, disapproval), and words that have similar grammatical functions such as conjunctions or adjectives (Brugidou & Le Quéau, 1999). These groups of related words are called *references*. A *relation* occurs when two references appear in the text in close proximity as part of a phrase or sentence (advanced readers can consult a user-friendly report on the Tropes® natural language processing features in Piolat & Bannour, 2009). The software also computes hierarchy and proximity graphs for the references identified, but these were not used as part of the current analysis.

We imported each of the transcribed text of all answers to each of the eight interview questions into the textual analysis software to generate eight independent reports. In order to extract cogent themes from the text analysis, we used the references generated by the textual analysis and the number of times these references co-occurred in the software output for a given interview question. We defined a theme as a conjunction of two references (i.e., relation) that appeared in the dataset three or more times. The combination of two references was intended to minimize the degree of inference required from raters to produce a meaningful theme from the resulting references. Specifically, a combination of two references provided sufficient context to produce a cogent theme with minimal assumptions on the part of the rater (e.g., references, support AND school → theme, "Special support needed at school"). This process was aided by additional context provided by the software output, which linked output references with portions of the

relevant input text. The integrity of this process was subsequently assessed with theme-generation inter-rater agreement analyses. We excluded themes that (a) were the result of references that could refer to multiple objects or actions (e.g., "thing" or "way"), (b) were composed of names of people as references (e.g., a therapist, or a particular child), and (c) were composed of references that were the result of the participant repeating a given word unnecessarily (e.g., participants on occasions repeated a word several times while gathering their thoughts). Table 2 presents a detailed technological description of the thematic analysis process (additional information is available in the Supplementary Online File, Appendices B and D).

2.5 | Inter-rater reliability

Some qualitative researchers question the importance of inter-rater reliability analysis, arguing that each code or theme is related to the unique perspective of the researcher. Thus, they argue that comparing the results from different researchers cannot inform us of the true nature of the phenomena being observed, and is instead just a perspective that two people acknowledge (Leung, 2015). However, the latter approach is at odds with the behavioral, conceptually systematic, and technological dimensions of applied behavior analysis. Therefore, it would still be important to produce technological descriptions and reliability analyses of item-generation methods, even if *qualitative* methods are involved (Yardley, 2000).

Unlike inter-observer agreement, inter-rater reliability in qualitative studies involves a second analyst replicating the methods of the research process in order to identify similarities and differences in the results produced (Morse, 2015). In the current analysis we used inter-rater agreement as a means to evaluate the extent to which different approaches to the analysis of the narrative record obtained during the interviews produced consistent themes. We obtained inter-rater agreement by having two raters re-code the narrative record from each interview according

TABLE 2 Textual analysis steps

Step	Description
1. Install and initiate Tropes® software and import data	Install latest software version from https://www.semantic-knowledge.com/download.htm go to File > Open to import the interview transcript to be analyzed.
2. Identify key reference relations	In the results tab, click on the 'references' box. Click on the first word appearing below the results tab and view the star graph produced. The word that is now in the middle of the graph is the central word, and will be referred to as this below. If the star graph shows words in blue or pink beside the central word with a number of three or more relations (small numeral next to the words related with the central word), this word and the central word potentially form a theme. Crosscheck the original text to discard irrelevant relations including the ones resulting from stuttering or unnecessary repetition, person-to-person relations, and meaningless noun-noun relations (e.g., way-thing).
3. Form themes	To determine the theme, click on the blue or pink word. In the 'extract' box above the star graph you will see all extracts of the text relating to this topic. Consider the relationship between these topics and the central word and describe this in one sentence, as a theme.
4. Continue the process	Continue steps #2 and #3 for all references and relations meeting the above criteria. Ignore relations that have already resulted in a theme previously.

to either the thematic analysis guidelines or the output of the textual analysis. We refer to these raters as *theme-generation raters* (raters #1 and #2). We then had two additional independent raters evaluate the number of agreements and disagreements in the themes identified by the primary and secondary theme-generation raters. We refer to these raters as *theme-similarity raters* (raters #3 and #4). We conducted the inter-rater agreement analysis both for the thematic analysis approach and the textual analysis approach. In order to avoid over reporting theme identification, a rater was involved only in the inter-rater agreement analysis of one approach (i.e., thematic or textual). The analysis was conducted on the basis of the answers to three of the eight open-ended interview questions produced by all participants (Q5, Q7, and Q8). Both analysts were postgraduate behavior analysis students. The instructions given to the raters are available from the Supplementary Online File (Appendix B).

The theme-similarity raters received the list of themes produced by both theme-generation raters and were instructed to draw arrows connecting the themes that matched, with the provisions that each theme could only be matched to one other theme, and that if a theme did not match another theme, they did not have to draw an arrow to connect that theme to any other theme (see an example of this process in the Supplementary Online Information, Appendix C). The same procedure was used with the themes resulting from both the thematic and textual analyses. We calculated a *theme-generation inter-rater agreement* (gIRA) as

$$\frac{\frac{A_{s_1} + A_{s_2}}{2}}{T_{g_1} + T_{g_2} - \frac{A_{s_1} + A_{s_2}}{2}} \times 100$$

where A_{s_1} is the number of themes identified by the primary theme-similarity rater as present among the themes produced by both theme-generation raters, A_{s_2} is the number of themes identified by the secondary theme-similarity rater as present among the themes produced by both theme-generation raters, T_{g_1} is the total number of themes produced by the primary theme-generation rater, and T_{g_2} is the total number of themes produced by the secondary theme-generation rater. The gIRA index varies from 0 to 100, lower values denote lower agreement in theme generation among theme-generation raters according to the mean judgment of the theme-similarity raters, whereas higher values denote higher agreement in theme generation among theme-generation raters according to the mean judgment of the theme-similarity raters. We expected the gIRA index to vary greatly across interview questions due to the possibility that the two methods of theme generation could lead to the generation of different (albeit equally relevant) themes.

Theme-similarity raters scored an agreement (A_s) when both theme-generation raters described a theme as referring to essentially the same topic. One theme-generation rater may have named a theme as *Service availability–families know that others are struggling, and talk about how long they waited to find this type of treatment since there is really nothing available in New Zealand*, whereas another may have used, *Alternatives–not many treatments available in New Zealand, especially one on one, intensive treatments*. Having essentially the same meaning, the theme-similarity rater would likely record an agreement when comparing these two themes. Thus, coding theme agreement required an abstract similarity judgment on the part of the theme-generation raters.

To determine the accuracy of the judgment of the theme-similarity raters that any two themes were the same or different we calculated a *theme similarity inter-rater agreement* (sIRA). An agreement occurred when both raters connected the same two themes. An agreement was also scored when both raters did not connect two themes (i.e., they agreed the themes were different). A disagreement occurred when one rater connected theme A to theme B but the second rater connected theme A to a theme different than B (or did not connect A to any theme at all). The sIRA was the result of dividing the number of agreements by the number of agreements plus disagreements and converting this ratio into a percentage.

2.6 | Social validity questionnaires

Questionnaires composed of Likert-scale items are often used to measure the social validity of interventions. Likert scales are more time efficient than qualitative methods (Gresham & Lopez, 1996). We generated two questionnaires based on thematic analysis and textual analysis, respectively. This resulted in a total of 23 items from the thematic analysis (Table 3) and 22 items from the textual analysis (Table 4). One theme from the thematic analysis and two themes from the textual analysis were not included as an item in the questionnaire since these statements would not be relevant to participants of other feeding programs, or would not provide any helpful information. These themes were: (a) being unsure of what to expect from treatment (thematic analysis), (b) concerns regarding the therapist being on holiday, and (c) a theme regarding the amount of meals provided, which was dependent on the family availability (textual analysis). A theme on funding and availability (thematic analysis) was split into two items, one on government funding and one on availability of treatment. Half of the resulting statements from each method were negatively phrased to reduce the effect of acquiescent response bias (Spector, 1992). In order to keep to limit the participants time commitment, parents were provided with a random selection of 23 questions each, with 12 questions from the thematic analysis and 11 questions from the textual analysis. Participants rated items over a 7-point Likert scale denoting their level of agreement with each statement (1 = strongly disagree; 7 = strongly agree). In order to inform content validity, participants responded yes or no to whether each item in the questionnaire was relevant, and whether it was easy to understand. We computed self-reported relevance and self-reported ease of understanding percentages

TABLE 3 Questionnaire items from the thematic analysis

1. Eating orally was not an obvious goal of the intervention.
2. The intervention had a negative impact on my family's stress levels.
3. The intervention was time consuming and stressful, which was not worth it for the end results.
4. The techniques that were a part of this intervention were not easy to use or effective.
5. This intervention was not acceptable from the standpoint of my cultural and religious beliefs.
6. Learning to eat did not open up any new possibilities to learn new skills for my child.
7. The support from and connection with the therapist/team was not positive.
8. This intervention should not be government funded.
9. The intervention did not allow my child to find likes and dislikes for different foods.
10. This intervention requires hard work and perseverance, and is not worth it for the end result.
11. This intervention did not give my child more freedom or confidence to have a social life.
12. It was not important for the child to remain the focus of the intervention.
13. It was good that the goals of intervention were increased slowly from smaller goals, to bigger, more challenging goals.
14. Support and communication, for example reminders and feedback, contributed to the ease with which I learned to use the techniques.
15. The benefits of this intervention outweighed any negatives involved.
16. The intervention aims and time constraints contributed to the family stress levels.
17. This intervention improved the day-to-day life of my child by teaching them structure and routine.
18. Without this intervention my family would be in a worse position.
19. The location where this intervention took place made it easier for me to continue with the intervention after the sessions with the therapist had finished.
20. Changing the mindset about food allowed my child to open up to new experiences.
21. The intervention opened up a whole new set of skills and lifestyle changes for my child.
22. There is nothing else available that is similar to this intervention to help children with feeding difficulties in my area.
23. Overall, this intervention used good strategies which made the hard work and emotional turmoil worthwhile.

TABLE 4 Questionnaire items from the textual analysis

1. I was not satisfied with the rate at which the level of challenge increased for my child during the intervention.
2. The intervention did not allow a progression toward eating more foods orally.
3. The techniques we were taught were not easy to use or highly effective.
4. I was not kept informed of what was happening during sessions.
5. Those providing the intervention did not believe that we would be successful.
6. This intervention did not reach beyond the scope of just eating.
7. Starting to eat again did not lead to enjoyment of food or better health.
8. Starting with small steps such as simple foods or small amounts did not make mealtimes easier.
9. It was not easy to keep up with all the data and information gathering required to do for the intervention.
10. Since the intervention had a limited duration (up to one year), it is less likely to have success than a longer intervention.
11. I had hoped to see my child eat a larger variety of foods by the end of the intervention.
12. The gradual approach to increasing oral eating was helpful.
13. The number of intervention sessions were enough to make progress.
14. To see progress with the problem we needed a new therapist/team to come in with a fresh perspective.
15. It was helpful that the therapist/team communicated with me and reminded me about what was happening.
16. Getting used to tasting new foods improved due to the intervention.
17. Drinking through the mouth was also included as an important part of the intervention.
18. This would be an important intervention for others in similar situations.
19. This intervention is beneficial provided other complications (eg. swallow problems) are ruled out first.
20. Each person involved in the intervention had a role to play in encouraging its success.
21. It did not take much time to learn the techniques before they became easy to use.
22. This intervention does not take a lot of time.

dividing the number of items for which the participant responded *yes* by the total number of items in the questionnaire, and converting that ratio into a percentage.

2.7 | Statistical analysis

Responses to the content validity questions were analyzed using an independent samples *t*-test to compare content validity of the thematic analysis items with the content validity of the textual analysis items. We used Levene's test to determine equality of variance between the two types of questionnaire. Equal variances were assumed for both understanding of the items and importance of the items since Levene's test was not significant.

In order to inform the validity of the social validity questionnaires we conducted known-group validity analyses with selected personal and treatment-related variables. This strategy can inform the validity of a newly created instrument by correlating the scale scores with predictors that are expected to differ across the levels of the construct of interest (i.e., social validity). Finding the expected relations provides an argument in favor of the scale's validity (Davidson, 2014). Ideally, the selection of variables should be guided by relevant literature.

While there is no specific evidence that client's age would moderate behavioral treatment acceptability (see e.g., Carter, 2007), the wide age range of participants in the current study, meant that variations in treatment perception due to client's age were likely. Second, Reimers et al. (1992) have reported that treatment duration may be

an important factor in the acceptability of behavioral interventions. Third, Hommel et al. (2013) encountered that treatment acceptability was a key factor in treatment success during a behavioral telehealth intervention. Finally, it is conceivable that social validity perceptions would shift as time elapses since the end of treatment (cf. memory bias, increased likelihood of relapse, long-term intervention effects). Therefore, we used the following predictors for our known-group validity analysis: age of the child in years at the beginning of treatment, the months the child spent in treatment in months, treatment success (whether tube feeding was discontinued, reduced, or did not change), and the time in months between the end of treatment and responding to the questionnaire.

We computed Spearman's rank correlation coefficients for several variables including the social validity questionnaire score (thematic analysis and textual analysis), the age of the child in years at the beginning of treatment, treatment success (whether tube feeding was discontinued, reduced, or did not change), the time the child spent in treatment in months, and the time in months between the end of treatment and completion of the questionnaire. We set the level of statistical significance at an alpha value of 0.05.

3 | RESULTS

The current study used both qualitative methods and quantitative methods. Hence, the results included both the quantitative results from questionnaires, as well as a short summary of the qualitative findings from interviews. The two methods for producing the questionnaires were also compared.

3.1 | Thematic analysis themes

For each interview question, between one and four themes were identified using thematic analysis by the primary analyst. The four themes described below were evident across several questions in the interview.

3.1.1 | Oral eating as a milestone

Participants identified the treatment leading to eating orally as having the additional benefit of opening up the child to new experiences. These new experiences allowed the development of new skills and lifestyle changes that previously they would not have been exposed to. Some of the experiences they describe include being able to go out into the community more since they do not have to bring all the equipment for tube feeding, going on camps and holidays, and learning new concepts such as routine and reward contingencies.

3.1.2 | Satisfaction-expectation trade-off

Dissatisfied families described being disappointed that the ultimate goal of becoming tube free was not met, while satisfied families described how they believed they would be in a worse position had they not received treatment. Although most families were unsure of what to expect when they began the treatment, families that tended toward higher expectations were less likely to describe themselves as satisfied. For example, one mother stated that she was not satisfied with the treatment because her child was not tube free; however, earlier in the interview, she reported that in hindsight, her expectations were too high.

3.1.3 | Benefits outweigh negatives

A common theme was that benefits of treatment outweighed any negatives that might be involved. Participants described how high costs in terms of time and stress were worth it for the end results. There was also a general consensus that although the treatment was demanding, the support from the therapist made it easier to cope in an otherwise challenging situation. Family members mentioned that seeing another person putting in an extraordinary amount of effort encouraged them to reciprocate. One mother summed up by saying "it's quite hard going sometimes but at the end of the day it's worth it because um..., you know, you show the results afterward of what happens."

3.1.4 | Funding and availability of treatment

All the participants referred to their struggle to find treatment for the child before being referred to the original study. Several of the participants had waited many years for help with tube weaning and getting the child to eat orally. There was an awareness that other families may be struggling with the same or similar problems, and have no access to treatment on the same level. Several participants expressed their desire for this to become a government funded treatment in New Zealand.

3.2 | Textual analysis themes

For the textual analysis, up to six themes were identified for each interview question by the primary analyst. The questions regarding religious and cultural beliefs did not generate any themes. This differed from the thematic analysis, where one theme was identified for the same question. There was less coherence across themes from different questions, with only two themes being evident across several questions in the interview.

3.2.1 | Gradual progress

Participants talked about treatment as a gradual succession from tube feeding to eating a variety of foods orally. Several participants mentioned how it was easier to start with small goals and gradually build up to more complex requirements. For example, one family mentioned that they did not think of using pureed food before, and instead had always offered foods they thought would be age-appropriate, which had not worked for them.

3.2.2 | Communication with the therapist

Communication and support from the therapist were regarded as key contributors to treatment success for some families. Reminders and clarification about the treatment plan and the way that this would play out each day was appreciated. Participants also liked to feel that they could be honest with the therapist about the treatment progress, and expressed that this helped them to negotiate achievable strategies to implement at home.

3.3 | Inter-analyst comparison

Table 5 (top panel) presents a summary of the consistency in theme generation and theme similarity evaluated by two independent raters. Themes produced by two separate analysts using the thematic analysis method differed consid-

TABLE 5 Theme generation inter-rater agreement (gIRA) and theme similarity inter-rater agreement (sIRA) across raters and methods

	No. themes generated		No. themes similar			
	Rater #1	Rater #2	Rater #3	Rater #4	gIRA (%)	sIRA (%)
Thematic						
Q5	2	5	2	0	17	80
Q7	3	5	3	2	45	80
Q8	3	6	1	1	13	89
Mean IRA					25	83
Textual						
Q5	2	3	2	2	67	100
Q7	3	2	2	2	67	100
Q8	3	4	3	2	56	75
Mean IRA					63	92
	No. themes generated		**No. themes similar**			
	Thematic	**Textual**	**Rater #3**	**Rater #4**	**gIRA (%)**	**sIRA (%)**
Method comparison						
Q1	5	4	2	1	20	90
Q2	6	3	1	3	29	89
Q3	0	1	0	0	0	100
Q4	2	4	1	1	20	100
Q5	2	2	0	0	0	100
Q6	3	3	2	2	50	78
Q7	3	3	1	2	33	89
Q8	3	3	0	1	9	89
Mean IRA					20	92

Note: Inter-method analysis based on Rater #1 theme counts. Q = interview question.

erably. Specifically, the mean gIRA was 25% (range, 13% to 45%) for interview items Q5, Q7, and Q8. Raters produced much higher sIRA values, with a mean agreement of 83% (range, 80% to 89%) for themes generated for interview items Q5, Q7, and Q8. By contrast, themes produced by two separate analysts using the textual analysis method produced relatively more consistent themes across raters. Specifically, the mean gIRA was 63% (range, 56% to 67%) for interview items Q5, Q7, and Q8. Raters produced high sIRA values, reaching a mean agreement of 92% (range, 75% to 100%) for themes generated for interview items Q5, Q7, and Q8.

3.4 | Inter-method comparison

Table 5 (lower panel) presents a summary of the consistency in theme generation and theme similarity produced by the primary raters using the thematic and textual methods. The primary rater obtained a mean gIRA value of 20% (range, 0% to 50%) across the two methods. The mean sIRA across all themes generated by the primary rater was 92% (range, 78% to 100%).

3.5 | Social validity questionnaires: Content and known-groups validity

We computed each participant average score on the social validity questionnaires generated by either thematic or textual analysis (Table 6). A higher score denoted a relatively higher level of self-reported satisfaction with the treatment. Overall participants agreed with statements produced from themes created using both methods. The mean item score across participants was 6.6 (standard deviation, 0.45) for the thematic analysis-based social validity questionnaire and 5.1 (0.68) for the textual analysis-based social validity questionnaire (score range, 1 to 7). Interestingly, the social validity questionnaire based on the thematic analysis produced slightly higher satisfaction scores across all participants.

The content validity analysis on the social validity questionnaires did not reveal significant differences between the two methods. Participants attributed the same relevance to both social validity questionnaires ($t_{10} = 0.95, p > 0.05$). Similarly, the ease of understanding of the items in the social validity questionnaires did not differ significantly across questionnaire development methods ($t_{10} = 1.08, p > 0.05$). In summary, 96% of the items derived from the thematic analysis and 91% of the items based on themes from the textual analysis were judged as *relevant*. Participants considered that 99% and 95% of items of the thematic and textual analysis, respectively, were easy to understand (Table 6).

As part of the known-groups validity analysis, Figure 1 presents the Spearman's rank correlation coefficients and regression lines between social validity score and average scores paired with child's age, treatment time, treatment success, and time to follow-up. There was a significant relation between the thematic analysis questionnaire score and the child's age ($r_s = 0.838, p = 0.019$), and treatment success ($r_s = 0.861, p = 0.014$). There was no significant relation between the textual analysis social validity questionnaire score and any of the predictors although this may be due to the limited statistical power of the analysis.

4 | DISCUSSION

In this study, we used semi-structured interviews with caregivers to inform the development of social validity questionnaires, a novel method for behavior-analytic research. Following interviews, we compared thematic analysis and textual analysis as two independent strategies to create a social validity questionnaire. The use of themes from open-ended interviews in this study provides a general approach to developing questionnaire items that is both technological and conceptually systematic and is not based on face validity alone. As mentioned previously, few scales have been developed for assessing social validity (e.g., TEI, IRP), and these lack detail on how items were generated, other than mentioning that they had face validity. Various methods have been proposed for generating and selecting

TABLE 6 Social validity scores and content validity analysis of questionnaires developed through thematic and textual analyses

	Social validity scores*		Self-reported relevance (%)		Self-reported ease of understanding (%)	
Participant	Thematic analysis	Textual analysis	Thematic analysis	Textual analysis	Thematic analysis	Textual analysis
P1	7.0 (0.00)	5.0 (2.79)	100	100	100	100
P2	6.5 (0.90)	5.5 (1.86)	91	100	100	100
P3	5.8 (1.82)	4.2 (2.27)	91	92	100	100
P4	6.9 (0.29)	4.4 (2.73)	92	82	100	100
P5	6.8 (0.45)	5.9 (1.20)	92	91	92	91
P6	6.5 (1.73)	5.5 (2.46)	100	91	100	82
Overall	6.6 (0.45)	5.1 (0.68)	96	91	99	95

Note: * Mean and standard deviation in parenthesis. Social validity score range, 1 to 7.

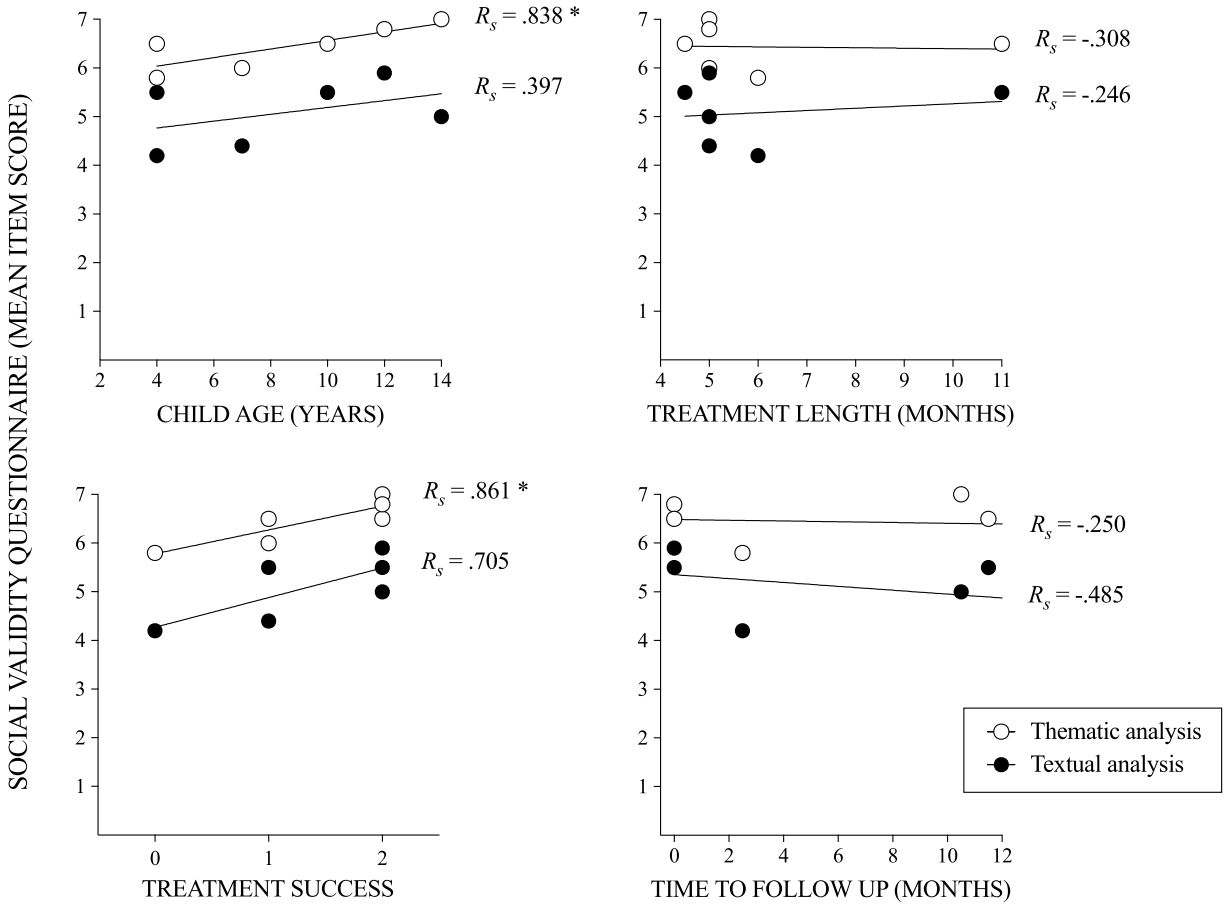

FIGURE 1 Correlation of social validity scores with child's age, treatment duration, treatment success, and time to follow up. *$p < 0.05$ (Spearman's rank correlation coefficients). Treatment success (bottom left panel) was coded as 0 (no change), 1 (tube feeding reduced), or 2 (tube feeding ceased)

items in the psychometric literature. However, these are typically intended for group-based applications and involve large number of test-takers for reliability and validity analyses. This study presents technological methods for developing social validity questionnaires for both individual and group-based applications, and presents some basic metrics to assess them (gIRA, sIRA).

The study also demonstrates the effective use of mixed methods to evaluate the social validity of a behavioral treatment program. Previously, Lyst et al. (2005) had used a combination of qualitative and quantitative methods to demonstrate the relevance of the Treatment Acceptability Rating Form-Revised (TARF-R). By comparing participants responses during interviews and focus groups to responses on the TARF-R, they were able to determine which factors of the TARF-R were also represented in the qualitative research. The current study extends this research by using qualitative methods to analyze interview responses that subsequently informed the development of a social validity questionnaire.

Overall, textual analysis was a more reliable method for generating themes that could be converted easily into a questionnaire composed of Likert-type items. While thematic analysis also provided themes that could be easily converted into Likert-type items, the inter-rater reliability of item generation (gIRA) was comparatively lower. Specifically, the themes produced by different analysts using textual analysis are more consistent than those produced by different analysts using thematic analysis. However, two independent raters also agreed that when themes produced using each method were compared, the themes often differed. Even though the interview transcriptions used to produce the themes were the same, employing different methods resulted in different themes being generated. In summary, thematic analysis produced very low levels of theme-generation reliability, whereas textual analysis produced moderate levels of theme-generation inter-rater reliability. Overall, both methods produced different themes. While all themes generated are arguably relevant, the agreement analysis suggests that those produced by the textual analysis approach are less likely to be a function of rater bias.

Baer et al. (1968) describe a "rule of thumb" for determining whether a procedure is technological, when provided with a written copy of the procedure, a trained reader should be able to replicate the results by using only the written instructions. In the current study, a trained reader was provided with written instructions on how to determine themes. The textual analysis approach reproduced the resulting themes to a higher degree than the thematic analysis. Replications are needed, preferably with diverse populations and behavior-analytic interventions, to determine the generality of these preliminary findings.

Interestingly, once the themes were converted to Likert-type items, a content validity analysis indicated that the thematic analysis produced a questionnaire containing more items that were deemed important to participants. It is unclear whether this difference was socially important. It is possible that textual analysis was more likely to produce themes that were not widely applicable across all participants. On occasions, a participant would mention that an item was not relevant to their particular situation and therefore rated it as *not important*. For example, a participant with a child that could not complete the intervention due to medical reasons rated *"Getting used to new foods improved due to the intervention"* as not important. The child had received initial treatment with an empty spoon and water only, so never had the chance to get used to new foods. In yet another example, two participants agreed that the statement "*This intervention was not acceptable from the standpoint of my cultural and religious beliefs"* was not important to them. Anecdotally, both participants stated that they did not adhere to any particular cultural or religious beliefs, but that this might be important to other people. Other participants rated this statement as important. Including items that are important to some individuals but not others supports the questionnaire's relevance and potential, but it may penalize its content validity as assessed in the current study. Future studies should add a "not applicable" response category so that questions that do not apply to a respondent's individual situation are not included in the overall score.

The thematic analysis items were rated as being marginally easier to understand than the textual analysis items, although this difference was not significant. A subsequent analysis showed that for both approaches negatively worded items were rated as more difficult to understand. Specifically, items that were not rated as easy to understand were all negatively worded (e.g., "Starting with small steps such as simple foods or small amounts did not make mealtimes easier"). Negatively worded items were included to reduce the participant's tendency to respond positively or neg-

atively in spite of the item content (Smith, 2004). However, participants found the negatively worded items in the questionnaires more confusing than other items. There is research to suggest that including negatively worded items can increase systematic error when the negative phrase is ignored and items are treated the same way as other items, which reduces the validity of the questionnaire (Podsakoff et al., 2003). Given these findings, we removed negatively worded items from the final questionnaire that is used for future recipients of the feeding intervention.

It is interesting that only the thematic analysis questionnaire scores were significantly correlated with personal and treatment variables in the differential analyses (known-group validity). Differential analyses are intended to provide an indirect means of validating a newly created metric. For example, Reimers et al. (1992) identified several variables that impact treatment acceptability ratings including reasonableness, effectiveness of treatment, side effects, cost of treatment, disruptiveness to family routines, and willingness to implement the treatment. Reimers et al. used the TARF-R to show that these variables did impact on acceptability ratings, even when the rating was completed at different follow-up appointments (1, 3, and 6 months). While the textual analysis did not produce statistically significant correlations, the general direction of the regression lines was consistent across both methods. Given that our correlation analyses were to some extent underpowered (only high-magnitude correlations would produce statistically significant coefficients), it is not possible to decide about the significance of these results. These analyses should be replicated with a larger sample of participants and with diverse outcome variables.

Parents and caregivers were chosen as social validity informants since not all the children receiving treatment could adequately respond to interview questions. All parents participating in the interviews were involved in implementing and observing treatment sessions at home. It would be useful to evaluate social validity from alternative perspectives including the children who received treatment and the health professionals involved in their care. While health professionals were not directly responsible for implementing treatment, they were highly involved in consultation and client referral. Extending social validity evaluation to include multiple stakeholders could provide a basis for an increased adoption of behavioral interventions.

Social validity measures were not recorded before the intervention began for any of the participants. Monitoring social validity before and after an intervention allows for the statistical analysis of social validity as an intervention outcome. In addition, pre-intervention social validity assessments allow for the inclusion of participants that will not complete the intervention (see a discussion of on-treatment vs. intention-to-treat analysis in behavioral interventions in Taylor et al., 2019). However, treatment-specific themes are often irrelevant before the intervention has begun (e.g., "Was the intervention difficult to implement?" "Was the intervention helpful?"). Therefore, pre-post social validity assessments may be restricted to spheres not involving the respondent perception of the treatment experience. For example, family burden and family stress are constructs that are relevant to social validity and may be relevant before, during, and after the intervention process (see e.g., Greer et al., 2008).

We recognize that thematic and textual analyses are not common in behavior-analytic research. However, these methods produced information that could not been obtained via quantitative methods. Participants described aspects of the intervention that extended beyond the narrower view of goals, processes and outcomes. For example, we identified themes that are not included on published surveys, such as benefits beyond eating, spiritual and cultural beliefs, and the importance of the provider demonstrating "belief" that the intervention would be successful. Further, our participants represented diverse cultural groups including New Zealand Māori, where qualitative methods promote "equal empowerment" (Barnes, 2000). Ultimately, the measurement of social validity should involve items that are meaningful to a diverse group of respondents.

There are some limitations to using transcripts of answers to open-ended interview questions as our source. First, there may be potential for acquiescence bias in participant responding (i.e., "saying what the researchers want to hear"; Anderson, 2010). In an attempt to minimize bias, we prevented the interviewer from being involved in the intervention. Moreover, the BCBA providing intervention was not involved in the interviews. A further limitation has been the lack of focus on reliability of the analysis, owing to the premise of the "unique" perspective of the researcher (Armstrong et al., 1997; Leung, 2015). Given our mixed-methods approach, we included agreement calculations to ascertain that textual analyses were a more reliable method. Thematic and textual analyses may help to identify themes

that are important to specific individuals only (e.g., *culture*, *spirituality*). In this connection, it may be possible to develop social validity questionnaires where respondents can choose to response to items that are most important to them. This approach may align with aspects of single-case design, such as the reporting of behaviors that are most socially significant to the individual and the aims of the intervention. Lastly, a qualitative approach may be time consuming, including the phases of interviewing, transcribing, and analyzing data (Braun & Clarke, 2006). A limitation of this study is that we did not record the time spent in these phases, and thus cannot make further statements regarding the time-efficiency of the process. The gIRA and sIRA metrics proposed here were intended to evaluate the proposed models for developing social validity questionnaires and account for a significant portion of the time cost of the procedures. However, we believe that these models can be replicated for individual and group applications without computing these metrics, which are primarily research-oriented.

4.1 | Future directions

The current line of research may be consolidated through systematic replications involving larger and diverse groups of respondents, client populations, and interventions. Replications are also needed to determine whether the proposed item-generation approaches are relevant both at the individual and group levels. Further studies may use textual analysis to assess the social validity of interventions for a specific problem or population (e.g., early intensive behavioral intervention, school-based intervention). Moving forward, questionnaires may then be produced that can be generally applicable within a population or intervention modality. Both of the questionnaires developed in this study are potentially usable for pediatric feeding interventions more widely. We note that the on-going use of social validity questionnaires developed through thematic and textual analyses would probably require additional reliability and validity analyses as it is customary within the confines of the classical test theory (see a summary of metrics and standards in Downing & Haladyna, 2006).

5 | CONCLUSIONS

This study compared two methods of analyzing interviews to produce a questionnaire for evaluating the social validity of behavioral treatments for feeding disorders. In addition, we proposed specific quantitative metrics to evaluate their reliability in the process of generating social validity themes (i.e., gIRA, sIRA). The textual analysis was a more reliable method of producing themes to be used as items within a social validity questionnaire and produced a questionnaire that had equal content validity to a questionnaire produced using themes from thematic analysis. Social validity assessments have often relied on questionnaires that were not developed for the specific population that they were being applied to. Commonly used questionnaires are often variations of existing scales which had been produced based on perceived face validity alone. Likewise, ad hoc social validity questionnaires rely solely on the judgment (perception of face validity) of the researcher. By contrast, the textual analysis method used in the current study provides a systematic approach to developing social validity questionnaires for specific populations that is minimally reliant on the researcher perception of face validity.

ACKNOWLEDGMENTS

ABA España provided support through a research contract with The University of Auckland (project no. CON02739). The current study was conducted in partial fulfillment of the requirements for the degree of Master of Science in Psychology awarded to Rachel Anderson at The University of Auckland. The authors express their gratitude to all study participants. Jessica C. McCormack, Sarah Anderson, Sarah Bushby, Maram Abumaree, Erin Zhai, and Veronika Rybová assisted with the data collection process.

CONFLICT OF INTEREST

The authors have no conflict of interest.

DATA AVAILABILITY STATEMENT

The data are not publicly available due to privacy or ethical restrictions. Part of the dataset is available upon request.

REFERENCES

Ahearn, W. H., Kerwin, M. L., Eicher, P. S., Shantz, J., & Swearingin, W. (1996). An alternating treatments comparison of two intensive interventions for food refusal. *Journal of Applied Behavior Analysis*, *29*(3), 321–332. https://doi.org/10.1901/jaba.1996.29-321

Anderson, C. (2010). Presenting and evaluating qualitative research. *American Journal of Pharmaceutical Education*, *74*(8), 141. https://doi.org/10.5688/aj7408141

Armstrong, D., Gosling, A., Weinman, J., & Marteau, T. (1997). The place of inter-rater reliability in qualitative research: An empirical study. *Sociology*, *31*(3), 597–606. https://doi.org/10.1177/0038038597031003015

Axe, J. B., & Sainato, D. M. (2010). Matrix training of preliteracy skills with preschoolers with autism. *Journal of Applied Behavior Analysis*, *43*(4), 635–652. https://doi.org/10.1901/jaba.2010.43-635

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, *1*(1), 91–97. https://doi.org/10.1901/jaba.1968.1-91

Bailey, J. S., & Burch, M. R. (2018). *Research methods in applied behavior analysis*. Routledge.

Barnes, H. M. (2000). Kaupapa maori: Explaining the ordinary. *Pacific Health Dialog*, *7*(1), 13–16. http://europepmc.org/abstract/MED/11709875

Berger, N. I., Manston, L., & Ingersoll, B. (2016). Establishing a scale for assessing the social validity of skill building interventions for young children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *46*(10), 3258–3269. https://doi.org/10.1007/s10803-016-2863-9

Borrero, C. S. W., Schlereth, G. J., Rubio, E. K., & Taylor, T. (2013). A comparison of two physical guidance procedures in the treatment of pediatric food refusal. *Behavioral Interventions*, *28*(4), 261–280. https://doi.org/10.1002/bin.1373

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brugidou, M., & Le Quéau, P. (1999). *Analysis of non-directive interviews with the «bundle» method*. World Association of Public Opinion Research Conference.

Bui, L. T. D., Moore, D. W., & Anderson, A. (2014). Using escape extinction and reinforcement to increase eating in a young child with autism. *Behaviour Change*, *30*(1), 48–55. https://doi.org/10.1017/bec.2013.5

Carr, J. E., Austin, J. L., Britton, L. N., Kellum, K. K., & Bailey, J. S. (1999). An assessment of social validity trends in applied behavior analysis. *Behavioral Interventions*, *14*(4), 223–231. https://doi.org/10.1002/(SICI)1099-078X(199910/12)14:4%3C223::AID-BIN37%3E3.0.CO;2-Y

Carter, S. L. (2007). Review of recent treatment acceptability research. *Education and Training in Developmental Disabilities*, *42*(3), 301–316. http://www.jstor.org/stable/23879624

Davidson, M. (2014). Known-groups validity. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and wellbeing research* (pp. 3481–3482). Springer. https://doi.org/10.1007/978-94-007-0753-5_1581

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Lawrence Earlbaum Associates.

Epstein, A., Whitehouse, A., Williams, K., Murphy, N., Leonard, H., Davis, E., Reddihough, D., & Downs, J. (2017). Parent-observed thematic data on quality of life in children with autism spectrum disorder. *Autism*, *23*(1), 71–80. https://doi.org/10.1177/1362361317722764

Ferguson, J. L., Cihon, J. H., Leaf, J. B., Van Meter, S. M., McEachin, J., & Leaf, R. (2019). Assessment of social validity trends in the journal of applied behavior analysis. *European Journal of Behavior Analysis*, *20*(1), 146–157. https://doi.org/10.1080/15021149.2018.1534771

Finn, C. A., & Sladeczek, I. E. (2001). Assessing the social validity of behavioral interventions: A review of treatment acceptability measures. *School Psychology Quarterly*, *16*(2), 176–206. https://doi.org/10.1521/scpq.16.2.176.18703

Foster, S. L., & Mash, E. J. (1999). Assessing social validity in clinical treatment research: Issues and procedures. *Journal of Consulting and Clinical Psychology*, *67*(3), 308–319. https://doi.org/10.1037/0022-006X.67.3.308

Fuqua, R. W., & Schwade, J. (1986). Social validation of applied behavioral research. In *Research methods in applied behavior analysis* (pp. 265–292). Springer. https://doi.org/10.1007/978-1-4684-8786-2_12

Galdas, P. (2017). Revisiting bias in qualitative research. *International Journal of Qualitative Methods*, *16*(1), 160940691774899. https://doi.org/10.1177/1609406917748992

Greer, A. J., Gulotta, C. S., Masler, E. A., & Laud, R. B. (2008). Caregiver stress and outcomes of children with pediatric feeding disorders treated in an intensive interdisciplinary program. *Journal of Pediatric Psychology*, *33*(6), 612–620. https://doi.org/10.1093/jpepsy/jsm116

Gresham, F. M., & Lopez, M. F. (1996). Social validation: A unifying concept for school-based consultation research and practice. *School Psychology Quarterly*, *11*(3), 204–227. https://doi.org/10.1037/h0088930

Hanley, G. P. (2010). Toward effective and preferred programming: A case for the objective measurement of social validity with recipients of behavior-change programs. *Behavior Analysis in Practice*, *3*(1), 13–21.

Hanley, G. P., Piazza, C. C., Fisher, W. W., Contrucci, S. A., & Maglieri, K. A. (1997). Evaluation of client preference for function-based treatment packages. *Journal of Applied Behavior Analysis*, *30*(3), 459–473. https://doi.org/10.1901/jaba.1997.30-459

Hanley, G. P., Piazza, C. C., Fisher, W. W., & Maglieri, K. A. (2005). On the effectiveness of and preference for punishment and extinction components of function-based interventions. *Journal of Applied Behavior Analysis*, *38*(1), 51–65. https://doi.org/10.1901/jaba.2005.6-04

Hawkins, R. P. (1991). Is social validity what we are interested in? Argument for a functional approach. *Journal of Applied Behavior Analysis*, *24*(2), 205–213. https://doi.org/10.1901/jaba.1991.24-205

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organisational Research Methods*, *1*(1), 104–121. https://doi.org/10.1177/109442819800100106

Hoch, T., Babbitt, R. L., Coe, D. A., Krell, D. M., & Hackbert, L. (1994). Contingency contacting: Combining positive reinforcement and escape extinction procedures to treat persistent food refusal. *Behavior Modification*, *18*(1), 106–128. https://doi.org/10.1177/01454455940181007

Hommel, K. A., Hente, E., Herzer, M., Ingerski, L. M., & Denson, L. A. (2013). Telehealth behavioral treatment for medication nonadherence: A pilot and feasibility study. *European Journal of Gastroenterology and Hepatology*, *25*(4), 469–473. https://doi.org/10.1097/MEG.0b013e32835c2a1b

Hunerberg, E. (2019). *3rd Eye theme analysis [online application]*. https://3rdeyeinformation.com/home

Kazdin, A. E. (1977). Assessing the clinical or applied importance of behavior change through social validation. *Behavior Modification*, *1*(4), 427–452. https://doi.org/10.1177/014544557714001

Kazdin, A. E. (1980). Acceptability of alternative treatments for deviant child behavior. *Journal of Applied Behavior Analysis*, *13*(2), 259–273. https://doi.org/10.1901/jaba.1980.13-259

Kelley, M. L., Heffer, R. W., Gresham, F. M., & Elliott, S. N. (1989). Development of a modified treatment evaluation inventory. *Journal of Psychopathology and Behavioral Assessment*, *11*(3), 235–247. https://doi.org/10.1007/BF00960495

Kennedy, C. H. (1992). Trends in the measurement of social validity. *The Behavior Analyst/MABA*, *15*(2), 147–156.

Kozlowski, A. M., Taylor, T., Pichardo, D., & Girolami, P. A. (2016). The impact of emerging liquid preference in the treatment of liquid refusal. *Journal of Developmental and Physical Disabilities*, *28*(3), 443–460. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=psyc13&AN=2016-14120-001

Kryzak, L. A., Bauer, S., Jones, E. A., & Sturmey, P. (2013). Increasing responding to others' joint attention directives using circumscribed interests. *Journal of Applied Behavior Analysis*, *46*(3), 674–679. https://doi.org/10.1002/jaba.73

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Leko, M. (2014). The value of qualitative methods in social validity research. *Remedial and Special Education*, *35*(5), 275–286. https://doi.org/10.1177/0741932514524002

Leung, L. (2015). Validity, reliability, and generalizability in qualitative research. *Journal of Family Medicine and Primary Care*, *4*(3), 324–327. https://doi.org/10.4103/2249-4863.161306

Lyst, A. M., Gabriel, S., O'Shaughnessy, T. E., Meyers, J., & Meyers, B. (2005). Social validity: Perceptions of check and connect with early literacy support. *Journal of School Psychology*, *43*(3), 197–218. https://doi.org/10.1016/j.jsp.2005.04.004

Meindl, J. N., Ivy, J. W., Glodowski, K. R., & Noordin, K. (2019). Applying standards of effectiveness to noncontingent reinforcement: A systematic literature review. *Behavior Modification*, *45*, 619–640. https://doi.org/10.1177/0145445519865073

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2010). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Morse, J. M. (2015). Critical analysis of strategies for determining rigor in qualitative inquiry. *Qualitative Health Research*, *25*(9), 1212–1222. https://doi.org/10.1177/1049732315588501

Piolat, A., & Bannour, R. (2009). An example of text analysis software (EMOTAIX-Tropes) use: The influence of anxiety on expressive writing. *Current Psychology Letters*, *25*(2), 1–23. https://doi.org/10.4000/cpl.4879

Podsakoff, P. M., MacKenzie, S. B., Lee, J., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Reimers, T. M., Wacker, D. P., Cooper, L. J., & DeRaad, A. O. (1992). Clinical evaluation of the variables associated with treatment acceptability and their relation to compliance. *Behavioral Disorders*, *18*(1), 67–76.

Rubio, E. K., McMahon, M. X. H., & Volkert, V. M. (2020). A systematic review of physical guidance procedures as an open-mouth prompt to increase acceptance for children with pediatric feeding disorders. *Journal of Applied Behavior Analysis*, *54*, 144–167. https://doi.org/10.1002/jaba.782

Rubio, E. K., Volkert, V. M., Farling, H., & Sharp, W. G. (2020). Evaluation of a finger prompt variation in the treatment of pediatric feeding disorders. *Journal of Applied Behavior Analysis*, *53*(2), 956–972. https://doi.org/10.1002/jaba.658

Schwartz, I. S., & Baer, D. M. (1991). Social validity assessments: Is current practice state of the art? *Journal of Applied Behavior Analysis*, *24*(2), 189–204. https://doi.org/10.1901/jaba.1991.24-189

Semantic Knowledge. (2014). *Tropes (version 8.4) [computer software]*. https://www.semantic-knowledge.com/company.htm

Sharp, W. G., Burrell, T. L., & Jaquess, D. L. (2014). The autism MEAL plan: A parent-training curriculum to manage eating aversions and low intake among children with autism. *Autism*, *18*(6), 712–722. https://doi.org/10.1177/1362361313489190

Sharp, W. G., Stubbs, K. H., Adams, H., Wells, B. M., Lesack, R. S., Criado, K. K., Simon, E. L., McCracken, C. E., West, L. L., & Scahill, L. D. (2016). Intensive, manual-based intervention for pediatric feeding disorders: Results from a randomized pilot trial. *Journal of Pediatric Gastroenterology and Nutrition*, *62*(4), 658–663. https://doi.org/10.1097/MPG.0000000000001043

Smith, P. B. (2004). Acquiescent response bias as an aspect of cultural communication style. *Journal of Cross-Cultural Psychology*, *35*(1), 50–61. https://doi.org/10.1177/0022022103260380

Spector, P. E. (1992). *Summated rating scale construction: An introduction*. Sage. https://doi.org/10.4135/9781412986038

Tarnowski, K. J., & Simonian, S. J. (1992). Assessing treatment acceptance: The abbreviated acceptability rating profile. *Journal of Behavior Therapy and Experimental Psychiatry*, *23*(2), 101–106. https://doi.org/10.1016/0005-7916(92)90007-6

Taylor, S. A., Purdy, S. C., Jackson, B., Phillips, K., & Virues-Ortega, J. (2019). Evaluation of a home-based behavioral treatment model for children with tube dependency. *Journal of Pediatric Psychology*, *44*(6), 656–668. https://doi.org/10.1093/jpepsy/jsz014

Taylor, T. (2020). Side deposit with regular texture food for clinical cases in-home. *Journal of Pediatric Psychology*, *45*(4), 399–410. https://doi.org/10.1093/jpepsy/jsaa004

Taylor, T., Blampied, N., & Roglić, N. (2020). Consecutive controlled case series demonstrates how parents can be trained to treat pediatric feeding disorders at home. *Acta Pediatrica*, *110*, 149–157. https://doi.org/10.1111/apa.15372

Thorne, S. (2000). Data analysis in qualitative research. *Evidence-Based Nursing*, *3*(3), 68–70. https://doi.org/10.1136/ebn.3.3.68

Ulloa, G., Borrero, C. S., & Borrero, J. C. (2019). Behavioral interventions for pediatric food refusal maintain effectiveness despite integrity degradation: A preliminary demonstration. *Behavior Modification*, *44*(5), 746–772. https://doi.org/10.1177/0145445519847626

Von Brock, M. B., & Elliott, S. N. (1987). Influence of treatment effectiveness information on the acceptability of classroom interventions. *Journal of School Psychology*, *25*(2), 131–144. https://doi.org/10.1016/0022-4405(87)90022-7

Walsh, R. S., McClean, B., Doyle, N., Ryan, S., Scarborough-Lang, S.-J., Rishton, A., & Dagnall, N. (2019). A thematic analysis investigating the impact of positive behavioral support training on the lives of service providers: "It makes you think differently". *Frontiers in Psychology*, *10*(2408). https://doi.org/10.3389/fpsyg.2019.02408

Wilder, D. A., Ertel, H. M., & Cymbal, D. J. (2020). A review of recent research on the manipulation of response effort in applied behavior analysis. *Behavior Modification*, *45*, 740–768. https://doi.org/10.1177/0145445520908509

Witt, J. C., & Martens, B. (1983). Assessing the acceptability of behavioral interventions used in classrooms. *Psychology in the Schools*, *20*(4), 510–517. https://doi.org/10.1002/1520-6807(198310)20:4%3C510::AID-PITS2310200420%3E3.0.CO;2-1

Wolf, M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis*, *11*(2), 203–214. https://doi.org/10.1901/jaba.1978.11-203

Wood, B. K., Wolery, M., & Kaiser, A. P. (2009). Treatment of food selectivity in a young child with autism. *Focus on Autism and Other Developmental Disabilities*, *24*(3), 169–177. https://doi.org/10.1177/1088357609338381

Woods, J. N., & Borrero, C. S. W. (2019). Examining extinction bursts in the treatment of pediatric food refusal. *Behavioral Interventions*, *34*(3), 307–322. https://doi.org/10.1002/bin.1672

Yardley, L. (2000). Dilemmas in qualitative health research. *Psychology and Health*, *15*(2), 215–228. https://doi.org/10.1080/08870440008400302

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Anderson, R., Taylor, S., Taylor, T., & Virues-Ortega, J. (2021). Thematic and textual analysis methods for developing social validity questionnaires in applied behavior analysis. *Behavioral Interventions*, 1–22. https://doi.org/10.1002/bin.1832