International Journal of Practice-based Learning in Health and Social Care
Vol. 3 No 1 2015, pages 58-76

An Investigation of the Standardised Patient Interview Rating Scale (SPIRS) for the Assessment of Speech Pathology Students in a Simulation Clinic

Anne E. Hill*
The University of Queensland, Australia

Bronwyn J. Davidson
The University of Melbourne, Australia

Deborah G. Theodoros
The University of Queensland, Australia

Abstract

Standardised patients (SPs) are increasingly utilised in health sciences education to assist students in the development of clinical competencies, including interviewing skills. This study investigated the development and validation of a rating scale for formative assessment of speech pathology students in an interview with an SP. Participants in this study were 76 undergraduate speech pathology students and 10 clinical educators who participated in a simulated clinic module. As part of the module, pairs of students interviewed an SP portraying a parent of a child with speech delay. The Standardised Patient Interview Rating Scale (SPIRS) was developed to assess students' foundation clinical competencies of communication, interviewing and professional practice skills. Students' interviews were videotaped, rated individually on the SPIRS by the clinical educator, and later re-rated by an expert rater. Data were analysed to determine the content validity, internal consistency, and inter-rater reliability of the tool. In addition, descriptive statistics were used to report student performance levels. Results indicated that the SPIRS had good content validity and internal consistency but that there may be some redundancy in individual items. An acceptable level of inter-rater reliability was achieved. Students generally scored highly, with non-verbal communication being the easiest and professional practice the most difficult skill to demonstrate. The SPIRS was found to be an appropriate tool for formative assessment of students in this simulated clinic module. Recommendations for improving its reliability were made. Further research is required to investigate use of the SPIRS as an assessment tool in other contexts utilising standardised patients.

Keywords: assessment; clinical education; speech pathology; standardised patients

Introduction

Clinical education programs for students in the health sciences increasingly include standardised patient experiences in addition to traditional placement opportunities, where students are typically supervised by a clinical educator in a workplace environment (Sheepway, Lincoln, and Togher 2011). The use of standardised patients (SPs) has been reported in studies investigating a variety of clinical learning situations and with a variety of participants, for example, assessing medical students in Objective Structured Clinical Examinations (OSCEs) (Hodges *et al.* 1998, Silverman *et al.* 2011), with undergraduate

*Corresponding Author: Dr Anne E. Hill, The University of Queensland, School of Health and Rehabilitation Sciences, Division of Speech Pathology, St Lucia QLD 4072, Australia
Email: ae.hill@uq.edu.au

physiotherapy students for history taking and physical examination (Ladyshewsky *et al.* 2000), with dietetics students to establish nutritional counselling skills (Henry 2007), and with international medical graduates seeking certification for practice through the United States Medical Licensing Examination (USMLE) (Clauser, Harik, and Margolis 2006). SPs are acknowledged to offer significant benefits to the teaching and learning situation: they provide an opportunity for students to practise their clinical skills in a safe environment prior to working with 'real' clients; their presentation can be manipulated to meet desired learning objectives; and they are able to give valuable feedback to students on their interview performance (Becker *et al.* 2006, Brigden and Dangerfield 2008, Hill, Davidson, and Theodoros 2010). While there has been a significant literature in medicine and nursing fields, reports of the use of standardised patients in speech pathology education programs have been limited (Hill, Davidson, and Theodoros 2010, Zraick 2012).

Students (or any 'examinee') participating in clinical tasks, such as those described above, require appropriate formative assessment and feedback to facilitate learning, and summative assessment to evaluate learning outcomes (Hattie and Timperley 2007, Norcini and Burch 2007). Such assessment is undertaken when the student is engaged in tasks typically associated with their profession (Crossley and Jolly 2012). It is therefore imperative that clinical assessment tools are developed with careful consideration of placement objectives, professional competencies expected, and ethical and valid assessment practices (McAllister *et al.* 2010, Petrusa *et al.* 1987).

The current study investigated formative assessment of speech pathology students completing an interview with an SP. Second year undergraduate students participated in a 12-week simulated clinic module, designed to develop their foundation clinical skills before entering into further clinical placements. Such experiences are considered important for students in the health sciences, to scaffold and facilitate their transition from the academic boundaries of the classroom to the real world environment of a workplace (Hill, Davidson, and Theodoros 2010, Le Maistre and Paré 2004). Students at early stages of their clinical learning are typically self rather than client-focussed, are nervous about interactions with clients and their families, and require significant direction and structure to manage clinical situations (McAllister and Lincoln 2004). Interviews are a typical starting point for assessment of a client and provide excellent opportunities to develop rapport, as a first step in an ongoing therapeutic relationship (Loftus and Mackey 2008). Providing interview opportunities for students allows them to apply the skills considered essential for speech pathology practice, such as interpersonal interaction, professionalism and communication skills, while supporting their learning and managing their anxiety. Furthermore, development of such generic and foundation skills provides a framework for the development of occupation-specific competencies, such as client assessment and diagnosis (McAllister *et al*. 2010).

The simulated clinic module incorporated three components: (1) three interviews with an SP portraying a parent of a child presenting with a speech disorder; (2) a series of clinical workshops involving discussion and practice in a range of clinical tasks (e.g., appropriate communication and interviewing); and (3) a kindergarten visit to practise administration and scoring of speech assessment tools. The SP interviews were considered a core component of the module. Students were assessed on their interview performance by clinical educators (CEs). A suitable tool was therefore needed for formative assessment of students' performance in interviewing a standardised patient.

A review of published single interview assessment tools revealed a number of instruments for consideration (Arthur 1999, McGraw and O'Connor 1999, Norcini *et al.* 2003, Silverman *et al*. 2011, Stillman, Sabers, and Redfield 1976, Stillman *et al.* 1977, Zraick, Allen, and Johnson 2003). Table 1 provides details of each of these tools. Within a speech pathology context, Zraick, Allen, and Johnson (2003) reported on a study investigating the use of SPs portraying adults with aphasia during OSCEs. The assessment tool published within their study was used by faculty staff to dichotomously rate speech pathology students ('done' or 'told' versus 'not done' or 'not told'), and contained seven items for assessment of non-technical skills, such as professional dress and maintenance of eye contact. The remainder of the tool centred on the student's ability to assess the SP's communication skills. Zraick, Allen, and

Johnson (2003) reported that inter-rater reliability evaluated on a subset of SP interviews was high.

Table 1. Tools to measure student performance in standardised patient (SP) interviews

Tool	Students assessed by	Features of rating scale	Internal consistency measures
Arizona Clinical Interview Rating Scale (ACIR Scale) (Stillman, Sabers and Redfield 1976; Stillman *et al*. 1977)	SPs ('trained mothers')	16 items on five point scale	Cronbach's alpha levels of 0.79 and 0.80 (two data sets)
(McGraw & O'Connor 1999)	SPs	19 dichotomous items	Not undertaken
Simulated Client Interview Rating Scale Arthur (SCIRS) (Arthur 1999)	SPs	39 items on three point scale	Cronbach's alpha levels greater than 0.90
Mini clinical evaluation exercise (mini-CEX) (Cook *et al*. 2010; Norcini *et al*. 2003)	Faculty staff	Six items plus overall rating on nine point scale	Cronbach's alpha levels of 0.86 to 0.88 (two data sets) (study by Cook *et al*. 2010)
EPSCALE (Silverman *et al*. 2011)	Hospital specialist, GP or communication specialist teacher	15 items on four point scale	Cronbach's alpha greater than 0.8

While the above tools were considered suitable for use in their reported clinical situations, a number of barriers to their use in the current study were identified. Firstly, the learning objectives in the current study were related not only to the development of foundation clinical skills but also to the practice of specific clinical skills for a case history interview within paediatric speech pathology practice. Whilst Zraick, Allen, and Johnson's tool (2003) addressed interpersonal, communication, and professional skills, these items were restricted to limited expected behaviours within a specific practice context of working with adults with aphasia. They were also rated dichotomously rather than on a scale. The other tools described above were situated within assessment of students in other professional environments, such as medicine and nursing.

Secondly, the majority of the other tools were developed for use by SPs in assessing students rather than for use by CEs, and primarily for summative assessment (Arthur 1999, McGraw and O'Connor 1999, Stillman, Sabers, and Redfield 1976, Stillman *et al.* 1977). Boulet *et al.* (2008) reported that performance in SP interviews for development of history-taking and physical examination skills is typically assessed by the SP. Indeed, much of the literature investigating assessment of students in SP interviews is focused on ratings by SPs, not by CEs (e.g., De Champlain *et al*. 1997, Huber *et al*. 2005, Tamblyn *et al*. 1991, Vu *et al*. 1992). Despite this, SPs are not always viewed as appropriate assessors of students by clinical staff (Vargas *et al*. 2007). Such resistance from clinical staff is possibly due to their unfamiliarity with SP use, concerns about SPs' capacity to adequately assess and provide feedback to students, and their fear that SPs might replace 'real patients' (Vargas *et al*. 2007).

Thirdly, CEs in the current study were required to rate the competencies of two students working as a pair within an interview of 20 to 25 minutes duration. This precluded lengthy

rating scales with many items, such as those reported above. A shorter scale which suited the conceptual framework and learning objectives of the simulated clinic module was required. Furthermore, in order to ensure that results of assessments undertaken using the tool were meaningful, its reliability, rater consistency and validity needed to be established (Downing 2004).

The aim of this study was to develop a tool suitable to assess speech pathology students' foundation clinical and specific professional skills during interviews with a standardised patient, in order to provide formative feedback. In addition, this study aimed to evaluate the content validity, internal consistency, and reliability of this tool.

Methods

Ethical clearance was obtained from the relevant ethics committee of an Australian university.

Development of interview assessment tool

Development of the interview assessment tool required consideration of four key factors: the content, the nature of the assessment (checklist or rating scale), the number of response options and the number of items to be included. Face and content validity were addressed initially through the process of item selection from a search of literature detailing clinical competencies and expert interview observation. Following the compilation of a list, all items were reviewed by experienced clinicians, to determine the suitability of their inclusion as relevant for an interview and representative of the underlying construct of foundation clinical competency. The final item list was a product of this process. As the simulated clinic module focused on the development of foundation clinical skills, the assessment tool was required to include items related to communication and interpersonal skills, interviewing skills, professionalism and specific clinical skills pertaining to a case history interview with a parent of a child presenting with speech delay. For example, items relating to the development of skills, such as rapport-building, were considered crucial.

A second consideration in the design of the tool was whether a rating scale or a dichotomous Yes/No checklist was indicated to assess the student's performance on each item. Rating scales are reported to be more vulnerable to discrepancy between raters than simple checklists (Tamblyn *et al*. 1991). However, rating scales are more widely used to evaluate interpersonal and communication skills. The less technical nature of these skills requires the tool to be sensitive to subtle differences in skill level, and to recognize the breadth of clinical competencies (Barry, Bradshaw, and Noonan 2013, Levine and Swartz 2008, Van Der Vleuten, Norman, and De Graaff 1991). In contrast, dichotomous checklists tend to be used in assessment of specific clinical examination techniques (Cohen *et al.* 1996). A rating scale was chosen as appropriate for assessing and providing feedback on clinical skills in the current study and, importantly, offered the opportunity to provide students with specific feedback on their performance level.

The number of available alternatives on the rating scale also required attention. Cox (1980) suggested that scales with two or three options only are usually inadequate in terms of their capacity to provide feedback. Conversely, too many options may lead to poor discrimination between rating points (Cook and Beckman 2009). An odd number of response alternatives is more effective than an even number so that a neutral position can be validly adopted. Cox (1980) concluded that five response alternatives was ideal, highlighting that such a number allowed not only for expected variation between examinees but also was effective in reducing the overuse of the 'neutral' category and allowing for valid ratings of examinees "along a continuum representing a single attribute" (1980: 409). Similarly, Cohen *et al.* (1996) suggested a 5 point scale for rating interpersonal and communication skills and used the terms: unacceptable; poor; average; good; and excellent. Understanding of each term can be facilitated by comprehensive instructions and labelling of the alternatives (Cox 1980) together with an acceptable level of training (Huber *et al*. 2005).

Finally, attention to the number of assessment items was considered important for maximising the accuracy of the instrument. Increasing the length of an assessment instrument may negatively affect agreement between raters, but the maximum acceptable number of items in an assessment is subject to debate (Vu *et al*. 1992). A wide range of between five and 30 items was proposed by Vu *et al*. (1992), with a suggestion that approximately 15 items may be considered appropriate for an assessment checklist completed within an interview. This was reiterated by Boulet *et al*. (2008) who suggested that limiting the item number rendered the list easier to create and to score, and potentially reduced costs associated with training.

Many clinical assessment rating scales include a global item, i.e. one which will provide a summation of a student's overall skills. Studies investigating the use of global judgements (Boursicot and Roberts 2006, Cunnington, Neville, and Norman 1997, Scheffer *et al*. 2008) have found that these may suit clinicians' understanding of clinical competency more than a list of items which attempt to separate discrete components of the competency. However, whilst these may be a time-efficient and appropriate method to assess communication and other professional skills, they may not provide the required level of detail to assist students in improving their performance (Scheffer *et al*. 2008). The inclusion of individual items, in addition to a global item, was deemed, therefore, to be a suitable choice.

The Standardised Patient Interview Rating Scale (SPIRS) – Case History form (Appendix 1) was formulated for the current study, in consideration of all of the above factors. Five categories of performance, on an ordinal scale from 1 to 5, were labelled as follows: unacceptable, poor, average, good, and excellent. A description of the performance at three of these categories provided anchor points for educator rating: these were 'unacceptable' (1), 'average' (3), and 'excellent' (5). The SPIRS included a total of six sections: non-verbal communication, verbal communication, interpersonal skills, interviewing skills, professional practice skills and specific clinical skills. Each of these six sections required a rating, and prompts for expected behaviours were included. For example, the verbal communication section included the following prompts - use of language and terminology, use of appropriate level of formality, speech volume, use of intonation, and speech rate. In addition, the tool included an 'overall rating of interview performance' and opportunity for specific written feedback from clinical educators, where required, on each of the sections.

Participants

Participants who consented to involvement in this study were 76 students in their second year of a four year undergraduate speech pathology program and 10 university-employed CEs with a mean of 18 years of clinical experience (range of six to 31 years). Students were aged between 19 and 39 years with a mean age of 20 years. There were two male and 74 female students.

Procedure

Student participants were introduced to the SPIRS in pre-clinical lectures at the start of their second year. Individual components of the tool, including the items and rating scale, were highlighted to increase their familiarity with the focus of formative assessment and confidence in using the tool. Students practised rating two contrasting videotaped case history interviews (one expert interview, one interview in which the clinician demonstrated unacceptable interview behaviours). Students were also assessed on a practice interview with an SP in a group context. Self-evaluation on the SPIRS also aided awareness. CEs were also familiarised with the SPIRS through group discussion of the tool features and practice in rating interviews, in order to ensure that the observations they made were accurate and appropriate, and that their written feedback to students was targeted and timely. SPs were provided with up to six hours of general communication and feedback training and an additional three hours of training which focused specifically on requirements for this simulated module, that is, how to provide appropriate verbal feedback to students, scenario content and practice in portrayal of the scenarios.

Students undertook the case history interview with the SP in week 5 of the 12-week simulated clinic module. Students worked in pairs but were assessed individually. A clinic-specific case history form, previously introduced to students, was used to guide questions during the interview. Each interview took approximately 20-30 minutes and was videotaped for future review. CEs used the SPIRS (Appendix 1) to rate students' performance during the case history interview in each of the six sections and, in addition, educators made an 'overall rating of interview performance' and provided specific written feedback to students on selected items. Where ratings were incomplete or CEs were equivocal, they viewed the videotaped interview to complete the rating. SPs also provided students with verbal feedback following their interview. The SPIRS was utilised as a formative assessment for each student, and provided evidence for CEs' ratings of students' clinical competency development at the end of the placement. Each of the 38 interviews (76 students working in pairs) was anonymised, then viewed and rated by an expert rater to determine inter-rater reliability. The expert rater (first author) was experienced in clinical competency assessment and in the use of the SPIRS to assess student performance in an SP interview.

Data analysis

The internal consistency of the SPIRS, as completed by each rater, was determined using Cronbach's alpha. A level of 0.7 or above was considered an indicator of a reliable tool (DeVellis 2003, Welkowitz, Cohen, and Ewen 2007). Internal consistency was further established using mean inter-item correlation (MIIC) measures with a value of between .2 and .5 considered acceptable (Briggs and Cheek 1986). Inter-rater reliability was analysed using percent exact agreement (PEA) and weighted kappas. These values were calculated on each item for each rater pair, for example, expert and rater 1 (total of 70 kappa values). Weighted kappas are suitable for evaluating inter-rater reliability on rating scale scores by taking into account chance agreement and allowing credit for partial agreement, where responses differ by one or two categories only (Streiner and Norman 2008). The strength of inter-rater reliability was determined using the guidelines reported by Landis and Koch (1977), that is, values <0 indicate poor agreement, 0.01 – 0.20 slight agreement, 0.21 – 0.40 fair agreement, 0.41 – 0.60 moderate agreement, 0.61 – 0.80 substantial agreement and 0.81 – 1.00 indicate almost perfect agreement.

Descriptive statistics were used to report the percentage of students who received ratings of 1 and 2, 3 and above, 4 and above, and 5 on each of the six items and the global rating score on the SPIRS. The Friedman test was used to compare student performance across items. Cronbach's alpha, inter-item correlation and Friedman test calculations were undertaken using the Statistical Package for the Social Sciences (SPSS) version 20. Weighted kappas were calculated using Stata Data Analysis and Statistical Software.

Results

Internal consistency measures are reported in Table 2. Cronbach's alpha levels were acceptable for the expert raters and all other raters with the exception of rater 6 (alpha level of .251). MIIC measures for two of 11 raters were in the range suggested by Briggs and Cheek (1986) with one measure below the 0.2 prescribed lower limit and the remainder above the upper limit of 0.5.

Table 2. Internal consistency measures for expert raters and each other rater

Rater	Cronbach's alpha level	Mean Inter-Item Correlations
Expert (n* = 76)	.926	.654
Rater 1 (n* = 12)	.837	.438
Rater 2 (n* = 11)	.838	.566
Rater 3 (n* = 6)	.957	.830
Rater 4 (n* = 5)	.804	.294
Rater 5 (n* = 6)	.914	.625
Rater 6 (n* = 4)	.251	.006

Rater 7 (n* = 8)	.888	.570
Rater 8 (n* = 12)	.894	.559
Rater 9 (n* = 6)	.925	.703
Rater 10 (n* = 6)	.962	.798

n* = number of students rated by this rater

Mean PEA across all seven items between the expert rater and each of raters 1 to 10 ranged from 71.57% to 92.86% with a mean of 82.06%. Kappa values ranged from 0.000 to 1.000. Individual PEA and kappa values for each item and each rater pair are detailed in Table 3 (see Appendix 1). Mean PEA for each item was above 70% with seven of 10 values above 80%. Forty one of 70 kappa values were in the substantial to almost perfect agreement range, with a further 15 values in the moderate agreement range. The lowest inter-rater reliability value, as measured by both PEA and kappa, was between the expert and rater 2, with the highest noted between the expert and rater 10. Items 3 (Interpersonal skills) and 4 (Interviewing skills) were the least reliable in rating, with lowest PEA and kappa values. PEA was highest for item 1 (Non-verbal communication), although similarly high PEA was achieved on items 2 (Verbal communication) and 5 (Professional practice skills). The highest reliability, with regard to kappa values, was also noted on item 2, with seven values (of 10) of almost perfect agreement and one of substantial agreement. The widest range of agreement was noted for items 7 (Overall rating) and 3.

Mean student performances on each item of the SPIRS, according to the expert rater and a mean of raters 1-10, are reported in Table 4 (see Appendix 1). A small proportion of students scored ratings of 1 or 2 on any item, while 96.05% or more scored 3 or above on all items. Raters scored 71.05% or more of students at levels 4 and above on all items and a range of 10.33 to 35.08% of students scored 5 on at least one item. The highest percentage of 5 ratings occurred for item 1, non-verbal communication, with the lowest on item 5, professional practice skills. The results of the Friedman Test indicated that there was no significant difference between students' performances on each of the items, according to the expert rater and the other raters.

Discussion

The SPIRS was found to have high content validity and internal consistency, and acceptable levels of inter-rater reliability, suggesting that it is an effective method of assessing students' clinical competency in case history interviews. This is an important finding given the increased attention on the inclusion of standardised patient experiences in clinical programs in speech pathology (Theodoros *et al.* 2010, Zraick 2012). The results of the current study compare favourably with other studies, reported in Table 1. Where internal consistency measures have been reported, these are noted to be 0.79 or more (Arthur 1999, Norcini *et al*. 2003, Silverman *et al.* 2011, Stillman, Sabers, and Redfield 1976, Stillman *et al*. 1977), a level achieved by all raters in the current study with the exception of rater 6.

The results of this study suggest that the SPIRS measures a unidimensional construct of clinical competence in an SP interview. Internal consistency scores for the expert and nine of 10 additional raters were high, indicating a correlation between all seven items on the SPIRS. Further investigation of the alpha result for rater 6 revealed that the score was diminished by ratings on items 1 and 2. Once these were removed from alpha calculations, the internal consistency score was at an acceptable level ($\alpha = 0.743$). Nine of 11 MIIC values fell outside of the guidelines defined by Briggs and Cheek (1986), with the majority of these being above the suggested level. Briggs and Cheek proposed that high MIIC values could reflect some item redundancy or over-specificity. Review of MIIC values for raters 4 to 8 revealed low inter item correlations (less than 0.2) between each of items 4 and 5 with item 1. A possible explanation for this may lie in the fact that item 1 was rated highly by all raters, achieving the

most scores of 5. In contrast, items 4 and 5 were least likely to achieve ratings of 4 or 5 on the ordinal scale. It would be appropriate to review all items to ensure that their inclusion in the scale is warranted and does not constitute unnecessary duplication.

Importantly, ratings on the global item (item 7) correlated highly with the other items. This suggests that the inclusion of a global item on the SPIRS was justified, a view confirmed by other studies (Cunnington, Neville, and Norman 1997, Scheffer *et al.* 2008). Crossley and Jolly (2012) noted that overall judgements may give a more reliable view.

An acceptable level of inter-rater reliability was achieved, with differing levels across items on the SPIRS and across expert and other rater pairs. In general, there was a good relationship between percent exact agreement and kappa values. The highest level of agreement was achieved for item 1 (Non-verbal communication). It could be posited that non-verbal communication is more transparently demonstrated and assessed, that is, it does not require interpretation, therefore allowing for higher agreement. In contrast, items 3 (Interpersonal skills) and 4 (Interviewing skills) achieved the lowest level of agreement, perhaps reflecting the requirement for interpretation of behaviours seen in the interview. The widest range of agreement level was noted for items 7 (Overall rating) and 3. This may also signal difficulties with interpretation of interview observations (in the case of item 3) and inconsistencies with 'summing up' a student's performance (item 7). Irrespective of the suggested reasons for reduced reliability, changes to procedure are required to minimise this.

Increased reliability may have been achieved through a number of considerations: improving the anchor descriptions for the rating scale, ensuring raters were clear on what was assessed within each item, and providing further training. Anchor descriptions provided on the SPIRS were generic in nature and limited to three of the five rating scale points. Providing more clinically-based descriptions of performance at each point of the scale, such as those provided by other authors (Iramaneerat *et al.* 2009, Panzarella and Manyon 2007, Silverman *et al*. 2011), may minimise differences in its interpretation (Crossley and Jolly 2012). Panzarella and Manyon (2007) reported, however, that there was only acceptable reliability on the four point rubric of their Integrated Standardized Patient Examination and suggested that improved rater training and practice trials would ameliorate some of these discrepancies. Training in the current study was limited to discussion of SPIRS features and practice in rating. More structured group rater training, with a focus on consideration of rating anchor points, may have prompted discussion needed to elucidate each of the rating points. For example, raters could re-rate students previously determined to be performing at levels of 'poor', 'average' and 'excellent', or script and role-play interviews at 'poor', 'average' and 'excellent' levels (Cook and Beckman 2009).

Crossley and Jolly (2012) suggested that raters can disagree on the interpretation of a rating scale, even when they agree on their observations. Furthermore, raters of a similar background can focus on different aspects of an interview they are watching, with contrasting results (Mazor *et al.* 2007). Mazor *et al*. (2007), in their study of assessment of medical students' professionalism in an OSCE situation, found that different raters evaluated professionalism by considering a range of contrasting behaviours, resulting in diverse rating outcomes. In the current study, shared discussion about both the scale and the observations therefore may be valuable. It is important to note, however, that even with optimal training structures in place, personal beliefs and practices, the impact of 'first impressions', and some level of subjectivity may lead to contrasting rater opinion (Crossley *et al*. 2002, Margolis *et al.* 2006, Wood 2014). Indeed, Margolis et *al*. (2006), in their generalisability study of ratings of examinees on the USMLE, found that 'rater stringency' had a more significant impact on rater variance than the nature of the task. For example, 'hawk' raters have a tendency to rate students lower than others, whereas 'dove' raters consistently rate highly compared with others (Crossley *et al*. 2002). In the current study, each of raters 1-10 rated different students, thereby limiting meaningful comparison of their rating stringency. Differences in agreement levels between the expert and each of raters 1-10 may be attributed to rater stringency but this is conjecture only and cannot be confirmed.

Group discussion would also resolve potential issues noted in the current study, related to attribution of observations to individual items. Review of SPIRS forms following quantitative analysis revealed that, in some cases of rater disagreement, raters had differed in ascribing a student's performance on one task to different items. Raters may have failed to distinguish between each of the six items on the scale (excluding the overall item), a behaviour also noted by Margolis *et al.* (2006). This is potentially confirmed by the MIIC values discussed above, wherein some items possibly duplicated others. Similarly, a 'halo effect' may result in a positive impression of one item impacting on others (Margolis *et al*. 2006), or an overall impression of a student's performance may impact on impressions of individual components (Clauser *et al.* 2012; Holmboe and Hawkins 1998). Presence of these rater behaviours is difficult to confirm but attention to their possible presence may improve reliability.

Reliability may have been affected by the fact that the expert rated 76 students, in comparison with each of the other raters who rated as few as four students and a maximum of 12. If indeed raters make judgments about ratings based on a 'yardstick' approach, wherein the performance of students may be compared with a previous student in completing a rating scale, then the expert rater had a more significant number of students with whom to make that comparison. In addition, the expert rater rated the students following video review, rather than at the time of their interview. The impact of this on decision-making regarding performance level is unknown, but it is possible that synchronous decision-making by raters 1-10 was influenced by other factors, such as more familiarity with the student and the SP and an understanding of the nature of previous discussions, such as the student's interview preparation.

Students generally achieved a high level of performance on the SPIRS. A small percentage of students achieved scores in the lower than average categories of 1 and 2 on the ordinal scale, with the majority of students scoring 3 and 4. The preponderance of 3 and 4 ratings indicates that students had gained interview skills through clinic discussion and practice, and that the interviews had provided students with the opportunity to demonstrate these skills. In addition, it should be acknowledged that some students were likely to enter the clinic with inherent or learnt skills in these areas. It is suggested that the frequent 3 and 4 ratings may be reflective of two trends. Firstly, raters tend to avoid extremes of scales when rating (Streiner and Norman 2008), and secondly, raters tend to be reluctant to give low ratings (Ginsburg, Regehr, and Mylopoulos 2009). It is possible, therefore, that the range of student performance was, in reality, more varied than these results suggest. A larger scale study of the SPIRS, with students from different year levels, would assist in determining its sensitivity to a broad range of skills.

As suggested above, non-verbal communication appeared to be the skill at which students performed best, perhaps due to its unambiguous nature and its common use in daily living. In contrast, professional practice skills were the most difficult for students, reflecting students' limited level of experience in the more profession-specific skill of case history taking and the higher requirement for integration of a range of skills. Analysis revealed, however, that although there were differences in the percentage of students achieving high ratings on each item, these differences were not statistically significant.

Limitations and future directions

A number of limitations were apparent in this study. Inter-rater reliability was examined by a single additional rater for each student rating at a different time and in a different medium. In addition, intra-rater reliability was not investigated. The SPIRS assessment was completed based on a student's performance in a single interview. Determining the clinical competence of a cohort of students from performance in one interview, particularly given a common tendency to perform differently when being viewed by others (Clauser *et al.* 2012), may not provide a valid view of their competence (Panzarella and Manyon 2007). In view of these factors, the use of the SPIRS as a means of providing formative assessment for students is

appropriate, but caution should be exercised in its use as a high-stakes summative assessment (Cook *et al*. 2010).

Items on the SPIRS form were developed by an expert group, based on a literature review and practical knowledge of required competencies. However, it may be that some important inclusions to the items were overlooked. An 'atomistic approach' (Clauser *et al.* 2012: 178) of identifying all separate and discrete items may lead to construct under-representation, where important components of skills under assessment are overlooked. Furthermore, the competence of a professional and their role cannot be easily defined by a set of individual items, so perhaps this is too simplistic (Crossley *et al.* 2002). The validity of the SPIRS would be further enhanced through its use in comparing the performance of two groups of students at different experience levels, to determine whether it is able to detect differences.

This study has identified a number of future directions for research. An evaluation of student learning outcomes, following formative feedback received from the SPIRS and students' perception of the value of the feedback, would be an important extension to this study. Although the SPIRS was designed for use with students in early simulation experiences, its content may be transferable to other contexts. As the SPIRS was found to have good content validity and internal consistency, it may have potential application for use as a summative assessment for health professional students in other simulated clinics, workplace settings, and OSCEs, as well as providing valuable feedback on students' performance in a clinical learning environment with a focus on formative feedback and learning. Additional modifications to meet the needs of alternative contexts would increase its utility, although further validation would then be required. An additional area of investigation is the potential for the SPIRS to be used by SPs, in their rating of students. As it has been noted that SPs rate students more leniently than experts (Scheffer *et al*. 2008), further attention to content and validation would be required.

Conclusion

Results of this study indicate that the SPIRS assessment of students' clinical competency in an SP interview has good content validity and internal consistency, but that there may be some redundancy in individual items. Acceptable levels of inter-rater reliability were achieved. As client interviews constitute an important component of speech pathology practice, it is likely that standardised patients will continue to be used to facilitate students' learning in this context. The validation of an appropriate interview assessment tool is, therefore, an important outcome. Given the inclusion of generic competencies, such as non-verbal and verbal communication and interpersonal skills, the SPIRS may have broader applicability and potential for use as an assessment tool for a range of interview contexts and student groups.

References

Arthur, D. (1999) 'Assessing Nursing Students' Basic Communication and Interviewing Skills: the Development and Testing of a Rating Scale'. *Journal of Advanced Nursing* 29, 658-665

Barry, M., Bradshaw, C. and Noonan, M. (2013) 'Improving the Content and Face Validity of OSCE Assessment Marking Criteria on an Undergraduate Midwifery Programme: A Quality Initiative'. *Nursing Education in Practice* 13, 477-480

Becker, K., Rose, L., Berg, J., Park, H. and Shatzer, J. (2006) 'The Teaching Effectiveness of Standardized Patients'. *Journal of Nursing Education* 45, 103-111

Boulet, J., Van Zanten, M., De Champlain, A., Hawkins, R. and Peitzman, S. (2008) 'Checklist Content on a Standardized Patient Assessment: an Ex Post Facto Review'. *Advances in Health Sciences Education* 13, 59-69

Boursicot, K. and Roberts, T. (2006) 'Setting Standards in a Professional Higher Education Course: defining the Concept of the Minimally Competent Student in Performance-Based Assessment at the Level of Graduation from Medical School'. *Higher Education Quarterly* 60, 74-90

Brigden, D. and Dangerfield, P. (2008) 'The Role of Simulation in Medical Education'. *The Clinical Teacher* 5, 167-170

Briggs, S. and Cheek, J. (1986) 'The Role of Factor Analysis in the Development and Evaluation of Personality Scales'. *Journal of Personality* 54 (1), 106-148

Clauser, B., Harik, P. and Margolis, M. (2006) 'A Multivariate Generalizability Analysis of Data from a Performance Assessment of Physicians' Clinical Skills'. *Journal of Educational Measurement* 43, 173-191

Clauser, B., Margolis, M., Holtman, M., Katsufrakis, P. and Hawkins, R. (2012) 'Validity Considerations in the Assessment of Professionalism'. *Advances in Health Sciences Education* 17, 165-181

Cohen, D., Colliver, J., Marcy, M., Fried, E. and Swartz, M. (1996) 'Psychometric Properties of a Standardized-Patient Checklist and Rating-Scale Form used to Assess Interpersonal and Communication Skills'. *Academic Medicine* 71, s87-s89

Cook, D. and Beckman, T. (2009) 'Does Scale Length matter? A Comparison of Nine- Versus Five-Point Rating Scales for the Mini-CEX'. *Advances in Health Sciences Education* 14, 655-664

Cook, D., Beckman, T., Mandrekar, J. and Pankratz, V. (2010) 'Internal Structure of Mini-CEX scores for Internal Medicine Residents: Factor Analysis and Generalizability'. *Advances in Health Sciences Education* 15, 633-645

Cox III, E. (1980) 'The Optimal Number of Response Alternatives for a Scale: a Review'. *Journal of Marketing Research* 17, 407-422

Crossley, J., Davies, H., Humphris, G. and Jolly, B. (2002) 'Generalisability: a Key to unlock Professional Assessment'. *Medical Education* 36, 972-978

Crossley, J. and Jolly, B. (2012) 'Making Sense of Work-Based Assessment: Ask the Right Questions, in the Right Way, about the Right Things, of the Right People'. *Medical Education* 46, 28-37

Cunnington, J., Neville, A. and Norman, G. (1997) 'The Risks of Thoroughness: Reliability and Validity of Global Ratings and Checklists in an OSCE'. *Advances in Health Science Education* 1, 227-233

De Champlain, A., Margolis, M., King, A. and Klass, D. (1997) 'Standardized Patients' Accuracy in recording Examinees' Behaviors using Checklists'. *Academic Medicine* 72, s85-s87

DeVellis, R. (2003) *Scale Development: Theory and Applications.* Thousand Oaks, California: Sage

Downing, S. (2004) 'Reliability: on the Reproducibility of Assessment Data'. *Medical Education* 38, 1006-1012

Ginsburg, S., Regehr, G. and Mylopoulos, M. (2009) 'From Behaviours to Attributions: Further Concerns regarding the Evaluation of Professionalism'. *Medical Education* 43, 414-425

Hattie, J. and Timperley, H. (2007) 'The Power of Feedback'. *Review of Educational Research* 77, 81-112

Henry, B. (2007) 'Use of the Standardized Patient Model to Develop Nutrition Counseling Skills'. *Journal of Nutrition Education and Behavior* 39, 50-51

Hill, A., Davidson, B. and Theodoros, D. (2010) 'A Review of Standardised Patients in Clinical Education: Implications for Speech-Language Pathology Programs'. *International Journal of Speech-Language Pathology* 12, 259-270

Hodges, B., Regehr, G., Hanson, M. and McNaughton, N. (1998) 'Validation of an Objective Structured Examination in Psychiatry'. *Academic Medicine* 73, 910-912

Holmboe, E. and Hawkins, R. (1998) 'Methods for Evaluating the Clinical Competence of Residents in Internal Medicine'. *Annals of Internal Medicine* 129, 42-48

Huber, P., Baroffio, A., Chamot, E., Herrmann, R., Nendaz, M. and Vu, N. (2005) 'Effects of Item and Rater Characteristics on Checklist Recording: What should we look for?' *Medical Education* 39, 852-858

Iramaneerat, C., Myford, C., Yudkowsky, R. and Lowenstein, T. (2009) 'Evaluating the Effectiveness of Rating Instruments for a Communication Skills Assessment of Medical Residents'. *Advances in Health Sciences Education* 14, 575-594

Ladyshewsky, R., Baker, R., Jones, M. and Nelson, L. (2000) 'Reliability and validity of an Extended Simulated Patient Case: a Tool for Evaluation and Research in Physiotherapy'. *Physiotherapy Theory and Practice* 16, 15-25

Landis, J. and Koch, G. (1977) 'The Measurement of Observer Agreement for Categorical Data'. *Biometrics* 33, 159-174

Le Maistre, C. and Paré, A. (2004) 'Learning in Two Communities: the Challenge for Universities and Workplaces'. *Journal of Workplace Learning* 16 (1/2), 44-52

Levine, A. and Swartz, M. (2008) 'Standardized Patients: the "Other" Simulation'. *Journal of Critical Care* 23, 179-184

Loftus, S. and Mackey, S. (2008) 'Interviewing Patients and Clients'. In *Communicating in the Health Sciences.* 2nd edn. ed. by Higgs,J., Ajjawi, R., McAllister,L. Trede, F. and Loftus, S.). South Melbourne, Victoria, Oxford University Press., 111-116

Margolis, M., Clauser, B., Cuddy, M., Ciccone, A., Mee, J., Harik, P. and Hawkins, R. (2006) 'Use of the Mini-Clinical Evaluation Exercise to rate Examinee Performance on a Multiple-Station Clinical Skills Examination: a Validity Study'. *Academic Medicine* 81, s56-s60

Mazor, K., Zanetti, M., Alper, E., Hatem, D., Barrett, S., Meterko, V., Gammon, W. and Pugnaire, M. (2007) 'Assessing Professionalism in the Context of an Objective Structured Clinical Examination: an In-Depth Study of the Rating Process'. *Medical Education* 41, 331-340

McAllister, L. and Lincoln, M. (2004) *Clinical Education in Speech-Language Pathology.* London: Whurr

McAllister, S., Lincoln, M., Ferguson, A. and McAllister, L. (2010) 'Issues in developing Valid Assessments of Speech Pathology Students' Performance in the Workplace'. *International Journal of Language and Communication Disorders* 45, 1-14

McGraw, R. and O'Connor, H. (1999) 'Standardized Patients in the Early Acquisition of Clinical Skills'. *Medical Education* 33, 572-578

Norcini, J., Blank, L., Duffy, F. and Fortna, G. (2003) 'The Mini-CEX: A Method for assessing Clinical Skills'. *Annals of Internal Medicine* 138, 476-483

Norcini, J. and Burch, V. (2007) 'Workplace-ABased Assessment as an Educational Tool: AMEE Guide No. 31'. *Medical Teacher* 29, 855-871

Panzarella, K. and Manyon, A. (2007) 'A Model for Integrated Assessment of Clinical Competence'. *Journal of Allied Health* 36, 157-164

Petrusa, E., Blackwell, T., Rogers, L., Saydjari, C., Parcel, S. and Guckian, J. (1987) 'An Objective Measure of Clinical Performance'. *The American Journal of Medicine* 83, 34-42

Scheffer, S., Muehlinghaus, I., Froehmel, A. and Ortwein, H. (2008) 'Assessing Students' Communication Skills: Validation of a Global Rating'. *Advances in Health Sciences Education* 13, 583-592

Sheepway, L., Lincoln, M. and Togher, L. (2011) 'An International Study of Clinical Education Practices in Speech-Language Pathology'. *International Journal of Speech-Language Pathology* 13, 174-185

Silverman, J., Archer, J., Gillard, S., Howells, R. and Benson, J. (2011) 'Initial Evaluation of EPSCALE, a Rating Scale that assesses the Process of Explanation and Planning in the Medical Interview'. *Patient Education and Counseling* 82, 89-93

Stillman, P., Brown, D., Redfield, D. and Sabers, D. (1977) 'Construct Validation of the Arizona Clinical Interview Rating Scale'. *Educational and Psychological Measurement* 37, 1031-1038

Stillman, P., Sabers, D. and Redfield, B. (1976) 'The Use of Paraprofessionals to teach Interviewing Skills'. *Pediatrics* 57, 769-774

Streiner, D. and Norman, G. (2008) *Health Measurement Scales: a Practical Guide to their Development and Use*. 4th edn. Oxford: Oxford University Press

Tamblyn, R., Klass, D., Schnabl, G. and Kopelow, M. (1991) 'Sources of Unreliability and Bias in Standardized-Patient Rating'. *Teaching and Learning in Medicine* 3, 74-85

Theodoros, D., Davidson, B., Hill, A. and MacBean, N. (2010) *Integration of Simulated Learning Environments into Speech Pathology Clinical Education Curricula: A National Approach* [on-line] available from https://www.hwa.gov.au/sites/uploads/sles-in-speech-pathology-curricula-201108.pdf [14 July 2014]

Van Der Vleuten, C., Norman, G. and De Graaff, E. (1991) 'Pitfalls in the Pursuit of Objectivity: Issues of Reliability'. *Medical Education* 25, 110-118

Vargas, A., Boulet, J., Errichetti, A., Van Zanten, M., Lopez, M. and Reta, A. (2007) 'Developing Performance-Based Medical School Assessment Programs in Resource-Limited Environments'. *Medical Teacher* 29, 192-198

Vu, N., Marcy, M., Colliver, J., Verhulst, S., Travis, T. and Barrows, H. (1992) 'Standardized (Simulated) Patients' Accuracy in recording Clinical Performance Check-List Items'. *Medical Education* 26, 99-104

Welkowitz, J., Cohen, B. and Ewen, R. (2007) *Introductory Statistics for the Behavioral Sciences.* New York: John Wiley and Sons Inc.

Wood, T. J. (2014) 'Exploring the Role of First Impressions in Rater-Based Assessments'. *Advances in Health Sciences Education* 19, 409-427

Zraick, R. (2012) 'Review of the Use of Standardized Patients in Speech-Language Pathology Clinical Education'. *International Journal of Therapy and Rehabilitation* 19, 112-118

Zraick, R., Allen, R. and Johnson, S. (2003) 'The Use of Standardized Patients to teach and test Interpersonal and Communication Skills with Students in Speech-Language Pathology'. *Advances in Health Sciences Education* 8, 237-248

Appendix 1

Standardised Patient Interview Rating Scale
Case History Interview

Student Name: _____ Year level: _____ Date: _____

Following the interview you have observed, please rate the student's performance by circling ONLY the appropriate number on the scale below. Student must be rated at one of the 5 points only (not in between) using the performance descriptors below as a guide. Listed below each skill area are some specific aspects to consider to help you rate the student's performance and formulate your feedback comments. You do not need to limit your comments to the specific aspects listed.

Unacceptable – Demonstrates many behaviours in specified skill area(s) that are inappropriate or have negative outcomes or consequences (make the situation worse). The desired outcome is not achieved.
Average – Demonstrates a sufficient range of expected behaviours in specified skill area(s) to achieve the desired outcome. Some deficiencies exist in the skill area(s) assessed but none are of major concern.
Excellent – Consistently demonstrates the full range of expected behaviours in specified skill area(s) to achieve the desired outcome. An outstanding level of performance is maintained. No deficiencies exist in the skill area(s) assessed.

1. COMMUNICATION / INTERPERSONAL SKILLS
The student demonstrated behaviours at the following performance levels in these skill areas:

Skills	Performance	Comments
1.1 Non-verbal Communication	unacceptable 1 poor 2 average 3 good 4 excellent 5	
	• eye contact • use of facial expression • body language • use of gesture	
	Performance	**Comments**
1.2 Verbal Communication	unacceptable 1 poor 2 average 3 good 4 excellent 5	
	• use of language and terminology • use of appropriate level of formality • speech volume • use of intonation • speech rate	
	Performance	**Comments**
1.3 Interpersonal Skills	unacceptable 1 poor 2 average 3 good 4 excellent 5	
	• building of rapport • response to client's feelings and needs • greeting of client (e.g. stood up, welcomed, introduced self, directed to seat) • allowing client to complete statements without interruption • maintenance of focus on client (e.g. note taking does not disrupt flow of interview)	

Skills	Performance	Comments
1.4 Interviewing Skills	unacceptable 1 poor 2 average 3 good 4 excellent 5	
	• use of open / closed questions / forced choice questions to gain specific information • use of verbal cues to indicate active listening • encouraging client to ask further questions • logical and systematic sequencing of questions • verification / clarification of information	
2.0 PROFESSIONAL PRACTICE SKILLS The student demonstrated the following performance on stated behaviours:		
Skills	**Performance**	**Comments**
2.1 Professional Practice	unacceptable 1 poor 2 average 3 good 4 excellent 5	
	• Explanation of professional role • Identification and understanding of the reason for visit • Summary of interview for parent • Discussion of follow-up plan • Interview conducted in a professional manner • Integration of knowledge, evidenced by what, when and how information is elicited during the interview • Respecting and maintaining professional boundaries	
3. CLINICAL SKILLS The student gained specific information in each history area below at the following performance level:		
	Performance	**Comments**
	unacceptable 1 poor 2 average 3 good 4 excellent 5	
(Please tick which history area this student completed)	• Birth/ developmental history □ • Speech and language history □ • Medical/family history □ • History of previous support □ • Educational history □ • Interaction and socialisation □	
4. OVERALL RATING OF INTERVIEW PERFORMANCE The student demonstrated communication, interview and interpersonal skills, professional and clinical practice skills at the following level during this interview:		
	unacceptable 1 poor 2 average 3 good 4 excellent 5	

Additional comments: _____

Clinical educator: _____ Student: _____

International Journal of Practice-based Learning in Health and Social Care
Vol. 3 No 1 2015, pages 58-76

Table 3. Agreement between expert rater and other raters on each item (weighted kappa and percent exact agreement)

Raters	Item 1 *Non-verbal communication*		Item 2 *Verbal communication*		Item 3 *Interpersonal skills*		Item 4 *Interviewing skills*		Item 5 *Professional practice skills*		Item 6 *Clinical skills*		Item 7 *Overall rating of interview performance*	
	PEA*	W kappa** values	PEA	W kappa values	PEA	W kappa values	PEA	W kappa values	PEA	W kappa values	PEA	W kappa values	PEA	W kappa values
Expert and rater 1	75	0.400	50	0.077	50	0.368	75	0.571	75	0.438	100	1.000	91.67	0.750
Expert and rater 2	90.9	0.667	100	1.000	54.55	0.444	55.56	0.000	75	0.500	50	0.000	75	0.000
Expert and rater 3	100	1.000	85.71	0.842	100	1.000	85.71	0.769	80	0.839	75	0.000	100	1.000
Expert and rater 4	100	nc§	100	1.000	100	1.000	75	0.667	100	1.000	75	0.000	25	0.000
Expert and rater 5	66.67	0.143	100	1.000	33.33	0.478	83.33	0.667	83.33	0.571	100	1.000	100	1.000
Expert and rater 6	75	0.000	75	0.500	75	0.500	50	0.000	100	1.000	100	1.000	100	1.000
Expert and rater 7	88.89	0.609	100	1.000	44.44	0.526	77.78	0.500	88.89	0.781	77.78	0.500	100	1.000
Expert and rater 8	91.67	0.869	91.67	0.906	58.33	0.712	58.33	0.546	100	1.000	58.33	0.531	83.33	0.733

**Corresponding Author:* Dr Anne E. Hill, The University of Queensland, School of Health and Rehabilitation Sciences, Division of Speech Pathology, St Lucia QLD 4072, Australia

Email: ae.hill@uq.edu.au

Expert and rater 9	100	nc[§]	85.71	0.842	100	1.000	85.71	0.800	100	1.000	71.43	0.647	71.43	0.600
Expert and rater 10	100	1.000	83.33	0.760	100	1.000	100	1.000	66.67	0.571	100	1.000	100	1.000
Mean across all raters	88.81	N/A	87.14	N/A	71.57	N/A	74.64	N/A	86.89	N/A	80.75	N/A	84.64	N/A

*Percent exact agreement.

**Weighted kappa.

[§]Not computed

Table 4. Percentage of students rating 1 and 2, 3 and above, 4 and above, and 5 on the ordinal scale on each item according to expert and other raters

Percentage of students with this rating	Ratings of 1 and 2		Ratings of 3 and above		Ratings of 4 and above		Ratings of 5	
	Expert rater	Mean of raters 1-10	Expert rater	Mean of raters 1-10	Expert rater	Mean of raters 1-10	Expert rater	Mean of raters 1-10
Item 1 *Non-verbal communication*	1.32	1.67	98.68	98.33	89.47	91.11	28.95	35.08
Item 2 *Verbal communication*	3.95	1.67	96.05	98.33	75	76.52	11.84	15.83
Item 3 *Interpersonal skills*	3.95	3.61	96.05	96.39	73.68	72.22	23.68	21.14
Item 4 *Interviewing skills*	1.32	0	98.68	100	71.05	68.89	10.53	12.5
Item 5 *Professional practice skills*	1.32	0	98.68	100	71.05	65.11	17.11	10.33
Item 6 *Clinical skills*	1.32	0	98.68	100	80.26	86.94	13.16	12.5
Item 7 *Overall rating of interview performance*	1.32	1.67	98.68	98.33	80.26	81.95	11.84	14.17