Reliability and validity of the Pragmatics Observational Measure (POM): A new observational measure of pragmatic language for children

Reinie Cordier [a,b,*], Natalie Munro [c], Sarah Wilkes-Gillan [c], Renée Speyer [b,d], Wendy M. Pearce [b]

[a] Curtin University, Perth, WA, Australia
[b] James Cook University, Townsville, QLD, Australia
[c] The University of Sydney, Sydney, NSW, Australia
[d] Department of Otorhinolaryngology and Head and Neck Surgery, Leiden University Medical Center, Leiden, The Netherlands

A R T I C L E   I N F O

A B S T R A C T

There is a need for a reliable and valid assessment of childhood pragmatic language skills during peer–peer interactions. This study aimed to evaluate the psychometric properties of a newly developed pragmatic assessment, the Pragmatic Observational Measure (POM). The psychometric properties of the POM were investigated from observational data of two studies – *study 1* involved 342 children aged 5–11 years (108 children with ADHD; 108 typically developing playmates; 126 children in the control group), and *study 2* involved 9 children with ADHD who attended a 7-week play-based intervention. The psychometric properties of the POM were determined based on the COnsensus-based Standards for the selection of health status Measurement INstruments (COSMIN) taxonomy of psychometric properties and definitions for health-related outcomes; the Pragmatic Protocol was used as the reference tool against which the POM was evaluated. The POM demonstrated sound psychometric properties in all the reliability, validity and interpretability criteria against which it was assessed. The findings showed that the POM is a reliable and valid measure of pragmatic language skills of children with ADHD between the age of 5 and 11 years and has clinical utility in identifying children with pragmatic language difficulty.

1. Introduction

This paper presents a psychometric evaluation of a new pragmatic assessment for school-aged children. Traditionally, pragmatics referred to the use of language in context (Prutting & Kirchner, 1987). The term related to verbal, paralinguistic and nonverbal aspects of communication during interactions with another. Pragmatics is a complex linguistic skill. For instance, how an adult explains, describes, asserts oneself and listens to another person, not only varies across communicative contexts and interactants, but also relies on how successful one is in engaging, maintaining and leaving that interaction. For school-aged children, important contexts and interactants include free play with peers (Cordier, Munro, Wilkes, & Docking, 2013). Play is an important occupation for children (Cordier & Bundy, 2009) and is a natural context for the development of pro-social behaviours and emotional regulation (Cordier, Bundy, Hocking, & Einfeld, 2010a). Peer–peer

* Corresponding author at: GPO Box U1987, Curtin University, WA 6845, Australia. Tel.: +61 404115786.
  E-mail address: reinie.cordier@curtin.edu.au (R. Cordier).

interaction and the social networks that follow also significantly increase from six to nine years of age (Feiring & Lewis, 1989). Indeed childhood social networks with peers (rather than adult-child interactions) continue to be important aspects of development and are critical during adolescence (Florey & Greene, 2008). This highlights the need to examine children's pragmatic skills within the context of play and with their peers.

Pragmatic skills are vitally important for successful peer–peer interaction and social-emotional development (Fujiki, Brinton, & Clarke, 2002; Jerome, Fujiki, Brinton, & James, 2002). Moreover, pragmatics, social skills, and emotional understanding are all interconnected (Hart, Fujiki, Brinton, & Hart, 2004; Russell & Grizzle, 2008). Contemporary approaches to pragmatics now define the construct as: "Behaviour that encompasses social, emotional, and communicative aspects of social language" (Adams, Baxendale, Llyod, & Aldredge, 2005, p. 568). Pragmatic assessment therefore warrants direct observational assessment in natural social contexts that also encompass behaviours across social, emotional and communicative domains of functioning (Adams, 2002; Leonard, Milich, & Lorch, 2011). This paper presents an analysis of the psychometric properties of an observational measure designed to assess the pragmatic language abilities of school-aged children as witnessed during peer–peer free play.

There are surprisingly few observational pragmatic language assessments appropriate for school age children. Instead, current practice involves parents, teachers and/or clinicians rating these behaviours across known communicative contexts and interactants using questionnaires (e.g., Children's Communication Checklist CCC-2; Bishop, 2003 and the Clinical Evaluation of Language Fundamentals – Fourth Edition Pragmatics Profile; Semel, Wiig, & Secord, 2003). Whilst these tools have established psychometric properties, directly observing peer–peer interaction not only allows for a deeper understanding about the nature of a child's current pragmatic skills, it also provides opportunity to develop naturalistic goals for psychosocial interventions when required (Gallagher, 2012). A recent systematic review of evidence-based treatment for school-age children with pragmatic language impairment identified only eight studies (from 1975 to 2008), all of which were described as exploratory (Gerber, Brice, Capone, Fujiki, & Timler, 2012). All eight studies used outcome measures developed by the investigators, five reported inter-rater reliability and only three used blinded raters. The authors of the systematic review noted that "…treatment on language use in social interaction is clearly in its infancy." (Gerber et al., 2012, p. 244). They also called for more research that examined "… the feasibility of outcome measures that document changes in children's language use across various partners (e.g., peers and teachers) and settings (e.g., classroom and playground)" (p. 246). Since the systematic review, only one treatment study has reported the use of an observational measure of pragmatics for school-aged children. In the randomised clinical trial of the social communication intervention project (Adams et al., 2012), the Targeted Observation of Pragmatics in Children's Conversation (TOPICC; Adams, Lockton, Gaile, & Freed, 2011) was used as a secondary outcome measure at 6 months after the intervention. Whilst this showed significant improvement in the intervention group relative to a treatment as usual group, the TOPICC's reliability and validity have yet to be reported. In addition, the TOPICC is a semi-structured task between an adult and a child that uses pictured photographs of events to elicit and evaluate the overall quality of a child's conversation skill. Clearly there is a need for a pragmatic language instrument that is broader in scope and also directly observes pragmatic skills of school-aged children during play interactions with a peer.

To our knowledge, there is only one direct observational pragmatic assessment of childhood pragmatic language with published psychometric properties. Developed by Prutting and Kirchner (1987), the Pragmatic Protocol (PP) was designed for adults and children aged 5 years and above. Prutting and Kirchner (1987) reported the inter-rater reliability to be .94. The PP asks clinicians to observe an adult–peer or child–peer interaction and evaluate verbal, paralinguistic and nonverbal aspects of communication. These are rated as appropriate, inappropriate or not observed. Assuming the clinician has had opportunity to observe the PP components, the rating scale then essentially becomes dichotomous and therefore does not capture a range of performances. A contemporary approach to pragmatics suggests that assessment of pragmatics should not be an all or nothing, appropriate-inappropriate phenomenon. Further, social and emotional aspects of peer–peer interaction are not included in the PP. In response to these limitations, a new observational assessment of pragmatics language was developed, named: Pragmatic Observational Measure (POM).

The overall aim of this study was to examine the psychometric properties of the POM. In order to create an observational instrument that could capture subtle performance across the range of behaviours that encompassed contemporary definitions of pragmatic language, a criterion design was utilised. Building on the work of Prutting and Kirchner (1987), Adams et al. (2005) and Brinton and Fujiki (2006), items were included that rated observed behaviours across the following five elements of pragmatic language: (1) introducing communication and being responsive to social interactions with peers; (2) interpreting and using non-verbal communication; (3) understanding and using emotional reactions and intentions of peers; (4) using higher-level thinking to promote interaction with peers; and (5) using appropriate negotiation techniques when interacting with peers.

2. Method

2.1. Participants

To examine the psychometric properties of the POM, data from two separate studies were included to enable investigation of *all* aspects of reliability and validity, including interpretability and responsiveness to clinical change. Data from *Study 1* (Cordier, Bundy, Hocking, & Einfeld, 2010b) included video footage of children with ADHD (*n* = 108), their

typically developing playmates (*n* = 108), and age-, sex-, and ethnic-matched typically developing children in a control group (*n* = 126). All 342 children were between the ages of 5 and 11 years. Children with ADHD were chosen because of their known social impairments and pragmatic language difficulties (Staikova, Gomes, Tartter, McCabe, & Halperin, 2013). The children diagnosed with ADHD were paired with playmates who were typically developing (one child with ADHD and one typical developing child in each observation); the control group consisted of typically developing children who were paired with a playmate who was also typically-developing (two typically developing children in each observation). All playmate pairs were familiar with each other. Children in the control group and the playmates of children with ADHD were known *not* to have ADHD, as defined by the DSM-IV criteria for ADHD (American Psychiatric Association, 2000).

2.1.1. Children with ADHD

This group included 108 children with ADHD recruited from district health boards and pediatricians' practices in Auckland, New Zealand. Diagnostic procedures were designed to ensure relatively high levels of diagnostic accuracy and to minimise the inclusion of borderline cases (i.e., cases just failing to reach criteria on the DSM-IV), and cases where diagnoses other than ADHD were deemed the primary diagnosis. To be included in the study, children had a formal diagnosis of ADHD made by a psychiatrist or paediatrician according to DSM-IV criteria. Additionally, children with ADHD were excluded if they were on medication where an overnight period was an insufficient wash-out (e.g., Atomoxetine). Parents/guardians were requested not to administer medication prescribed for ADHD on the day of the assessment to enable observation of how children with ADHD interact without the effects of medication.

2.1.2. Typically developing children in the control group

The 126 participants were recruited from professional networks such as local schools and from families of health service employees in Auckland, New Zealand. For the purpose of this paper, a typically developing child was defined as a child who did not have a childhood developmental disorder (i.e., scored below the clinical cut-off for any of the Conners' Parent Rating Scales-Revised [CPRS-R] subscales and DSM-IV scales) and for whom no concerns had been raised about development by a teacher or health professional. The CPRS-R was administered for all children in the sample. The CPRS-R is a paper and pencil screening questionnaire completed by parents/primary carers to assist in determining whether children between the ages of 3 and 17 years have signs and symptoms consistent with the diagnosis of ADHD. The CPRS-R has evidence of excellent reliability (international consistency reliability 0.75–0.94) and construct validity (to discriminate ADHD from the non-clinical group: sensitivity 92%, specificity 91%, positive predictive power 94%, negative predictive power 92%) (Conners, 2004; Conners, Sitarenios, Parker, & Epstein, 1998). The demographic information from the participants and their primary caregivers and the CPRS-R subscale scores are summarised in Table 1.

Children from *Study 2* (Cordier, Ling, Munro, & Wilkes-Gillan, in preparation) involved nine children with ADHD who participated in a seven-week play-based intervention. The children included eight boys and one girl (mean age = 8.2 years; 1.1 SD) with a formal and unambiguous diagnosis of ADHD made by a psychiatrist or paediatrician using recognised diagnostic procedures (i.e., meeting the criteria for ADHD as defined in the Diagnostic and Statistical Manual of Mental Disorders, 4th edition [DSM-IV]). The presence of ADHD symptoms was confirmed a priori by parent ratings on the CBRS-P.

2.2. Instruments

2.2.1. Pragmatic Observational Measure (POM)

The POM was developed out of the need for an observational measure to assess pragmatic language in naturalistic contexts during peer–peer interactions. Only two observational measures were found to measure some aspects of pragmatic language: the PP (Prutting & Kirchner, 1987) and the Structured Multidimensional Assessment Profiles (S-MAPs; Wiig, Larson, & Olson, 2004). S-MAPs are rubric assessments that examine many dimensions of spoken and written language via observation and/or portfolio examination. For a particular dimension, a child's level of performance is rated from beginner to expert (along a scale of 1–4) across dimensions. Whilst the S-MAPs was not specifically developed to measure pragmatic language, it did contain rubrics related to some aspects of pragmatic language. The S-MAPs was developed as a tool for clinicians for curriculum-based assessment and intervention for grades K-12 with examples of dimensions and items and a description for clinicians on how to develop their own rubrics. However, the S-MAPs has not been assessed for its psychometric properties, which restricts it usefulness within a research context. Similarly, the PP has limitations; only limited information on the psychometric properties has been published and the dichotomous appropriate versus inappropriate scale is problematic, given the complex and multifaceted nature of pragmatic language.

Therefore, the aim of the POM is to measure the pragmatic language performance of children between the ages of 5 and 11 years during peer–peer interactions within naturalistic contexts. Using the contemporary definition of pragmatic language, the POM was designed to map the five elements of pragmatic language: (1) *Introduction and responsiveness* (introducing communication and being responsive to social interactions with peers); (2) *Non-verbal communication* (interpreting and using non-verbal communication); (3) *Social-emotional attunement* (understanding and using emotional reactions and intentions of peers); (4) *Executive function* (using higher-level thinking to promote interaction with peers); and (5) *Negotiation* (using appropriate negotiation techniques when interacting with peers). Table 2 provides a summary of the five elements and a summative description of each item that are grouped within each of the five elements.

Table 1
Participant demographics.

Child and parent demographics	Typically developing children (n = 126)	Children with ADHD (n = 108)	Playmates of children with ADHD (n = 108)
Mean age (SD)	8.6 years (1.5)	8.9 years (1.4)	8.4 (1.9)
Percentage boys vs. girls	78.7%/21.3%	80.3%/19.7%	75.2%/24.8%
Ethnicity			
European	65.2%[a]	67.8%[a]	65.2%[a]
Maori	19.7%[a]	16.1%[a]	16.1%[a]
Other ethnicities	15.1%[a]	16.1%[a]	18.7%[a]
CPRS-R subscale scores			
Oppositional	50.6[b]	70.4[c]	56.9[b]
Cognitive problems	49.5[b]	72.5[c]	51.4[b]
Anxious/shy	50.8[b]	58.9	50.3[b]
Perfectionism	49.3[b]	56.1	49.8[b]
Social problems	48.9[b]	76.0[c]	60.4[b]
Psycho-somatic	50.6[b]	64.4	49.8[b]
Emotional labile	48.5[b]	62.8	50.6[b]
Behavioural problems	49.7[b]	73.0[c]	56.2[b]
Primary carer's highest level of education			
Did not complete high school	19.1%	13.4%	10.7%
Completed high school	46.8%	40.2%	39.3%
Completed tertiary qualifications	34.1%	46.4%	50.0%
Primary carer's occupation			
Jobs that do not require tertiary qualifications	75.4%	63.4%	58.9%
Jobs that do require tertiary qualification	24.6%	36.6%	41.1%

[a] This is a close approximation of the current ethnic distribution of the New Zealand population estimate (Statistics New Zealand, 2007) with Europeans making up 76.8%, Maori 14.9% and the remainder of ethnic groups 17.8% of the population, thus representative of the New Zealand population.
[b] CPRS-R subscale mean scores are all below the clinical cut-off (i.e., subscale scores > 70).
[c] Denotes CPRS-R subscale mean scores above the clinical cut-off (i.e., subscale scores > 70).

Table 2
Pragmatic Observational Measure element and item description.

POM items	Summative item description
Element: Introduction and responsiveness	
1. Select and introduce	Selects and introduces a range of conversational topics
2. Maintain and change	Maintains and changes conversational topics appropriately
3. Contingency	Shares or adds information to the previously communicated content
4. Initiate	Initiates verbal communication appropriate to the context
5. Respond	Responds to communication given by another
6. Repair and review	Repairs and reviews conversation when a breakdown in communication occurs
Element: Non-verbal communication	
7. Facial expression	Uses and responds to a variety of facial expressions to express consistent meanings
8. Gestures	Uses and responds to identifiable, clear, intentional body actions or movements
9. Body posture	Uses and responds to clear, identifiable body positioning and stance
10. Distance	Use of physical space between speakers
Element: Social-emotional attunement[a]	
11. Emotional attunement	Being aware of and responsive to another's emotional needs
12. Self-regulation	Regulate own thinking, emotions and behaviours
13. Perspective taking	Considers/integrates another's viewpoint/emotion
14. Integrating communicative aspects	Appropriate use of social language within context
15. Environmental demands	Adapts behaviour to environmental demands
Element: Executive function	
16. Attention, planning, initiation	Attends to communicative content, plans and initiates appropriate responses
17. Communication content	Interprets, plans, organises and delivers content
18. Creativity	Versatile ways to interpret/connect/express ideas
19. Thinking style	Thinks and articulates abstract and complex ideas
Element: Negotiation	
20. Conflict resolution	Uses appropriate methods for resolving disagreement
21. Cooperation	Works together; mutually beneficial exchange
22. Engagement/interaction	Consistently gets along well with another peer whilst engaged
23. Assertion	Makes clear own opinions, viewpoints and emotions
24. Express feelings	Expresses feelings appropriate to the context
25. Suggests	Makes suggestions and offers opinions
26. Disagrees	Disagrees in an effective way that promotes the interaction
27. Requests	Requests explanations/more information in an effective way

[a] The item discourse interruption (originally included under the element social-emotional attunement) was removed following factor analysis.

Each of the 27 items is rated on a 4 point scale (1–4) based on the child's consistency of performance ranging between: 1 – rarely or never observed; 2 – sometimes observed (25–50% of the time); 3 – much of the time observed (50–75% of the time); and 4 – almost always observed (75–100% of the time). A detailed description is provided for each level of performance for all items. The items were selected, developed and refined by the first three authors over a period of 18 months. All the authors have extensive experience in working with children from four disciplines, clinical psychology, epidemiology, speech and language pathology and occupational therapy. The item level descriptors were continuously refined over this period to ensure that they were clear, unambiguous and that all items could be rated using observable behaviour. Item refinement also involved external raters who rated video footage of children who were typically developing as well as children who had behavioural disorders, and pragmatic language difficulties. Feedback from these raters was used to further refine item descriptors.

2.2.2. Pragmatic Protocol (PP)

As the PP is the only observational assessment with reported psychometric properties, this instrument was used as the reference tool (i.e., "gold standard") against which the POM was evaluated. The PP consists of seven components of pragmatic language which comprise 30 corresponding items and is designed for observational assessment of school-aged children and adults. The authors recommended that the PP should be completed after observing an individual during 15 minutes of "spontaneous, unstructured, conversation with a communicative partner" (Prutting & Kirchner, 1987, p. 108). The rater then indicates whether a child's use of the pragmatic parameter under question was appropriate or inappropriate. In the current study, three PP components were rated: Verbal aspects-topic (VAT), Turn taking (TT), and Nonverbal aspects (NVA) as these fit the contemporary definition of pragmatic language that was adopted.

2.3. Procedure

Ethical approval was obtained from the University of Sydney Human Ethics Research Committee to perform secondary analysis on data from *Study 1*. The original study aimed to compare the play skills of children with ADHD with typically developing children (Cordier et al., 2010b). In *Study 1*, peer–peer social interactions for children in the control group were observed, using a designated play area at the respective schools that children attended, and for children with ADHD at clinics that they regularly attended. Toy selection catered to age and gender differences and a diversity of play materials was present in each room to support a range of play in order to encourage peer–peer interaction. The same toys were present during all play sessions and the children were allowed to choose their play materials and activities.

The assessor introduced the peers to the free play situation. Participants were instructed that they could play with any of the toys in the playroom for 20 minutes and that they should ignore the assessor who was present in the playroom. The assessor was as unobtrusive as possible and did not intervene unless a child was in danger. When children attempted to interact with the assessor, the assessor's response was neutral. The play session was videotaped for later analysis.

Study 2 involved 9 children with ADHD and their typically developing playmates attending weekly play sessions for the duration of seven weeks (Cordier et al., in preparation). The play-based intervention involved: parent training, three clinic play-sessions, weekly home-based modules (interactive DVD/manual) and therapist phone consultations. The clinic play-sessions involved the use of video-feedback and -feed forward, therapist modelling and parent training aimed at developing the play, social and pragmatic skills of children with ADHD. Video footage from for all 9 children with ADHD pre-post intervention was used to examine the interpretability of the POM.

A single experienced rater (who was not the assessor) rated all the children from the videotapes from both *Study 1 and 2*. The rater was blinded to the purpose of the current study to minimise bias. To establish adequate inter-rater reliability, the following procedures were employed. First, another blinded rater familiarised herself with both instruments (PP and POM). Next, the first and third author developed a training video using footage of school-aged children playing who were independent of the current study. Ten video samples were viewed and independently coded by the blinded rater and the third author using the PP and POM. Coding was compared and then consensus reached following discussion and re-viewing of the training footage. Reliability for the current study was calculated based on a random selection of 30% of all data.

2.4. Statistical analysis

The COSMIN taxonomy of psychometric properties and definitions for health-related outcomes (Mokkink et al., 2010a, 2010b; Terwee et al., 2012) were used to investigate the psychometric properties of the POM. The POM scores of the typically developing children from *Study 1* ($n = 126$) were used to examine the following psychometric properties: internal consistency, reliability, measurement error, content validity, construct validity, and criterion validity. Interpretability is not considered a psychometric property but an important characteristic of an instrument and was determined to assign qualitative meaning to quantitative data (Mokkink et al., 2010a). The data from children with ADHD ($n = 108$) and typically developing children in the control group ($n = 126$) in *Study 1* were used to examine responsiveness and another component of construct validity. To examine interpretability, the data from all children in *Study 1* (ADHD = 108; playmates = 108; typically developing children in the control group = 126; $N = 342$), as well as the 9 children from *Study 2* were used.

2.4.1. Reliability

2.4.1.1. Internal consistency. Internal consistency is a measure of the extent to which items in a questionnaire subscale are correlated, thus measuring the same concept (Streiner, 2003). As a first step, correlations between items that relate to a particular element were calculated using Spearman's Rho (r_s) to determine if they belonged to the same construct being measured. Thereafter an exploratory principal component (PC) factor analysis was performed to determine if the POM measured a uni-dimensional or multi-dimensional construct. This was followed by a Maximum Likelihood (ML) factor analysis once the likely number of factors was determined. Given the relatively small sample size, further analysis was performed to determine internal-consistency reliability. The internal consistency reliability was examined by calculating the Cronbach's α coefficient. A low Cronbach's α (<0.70) suggests a lack of correlation, whereas a high Cronbach's alpha (>0.90) indicates redundancy of one or more items.

2.4.1.2. Reliability. Reliability refers to the degree to which participants can be distinguished from each other, despite measurement error (Streiner & Norman, 2003). Reliability was assessed by calculating the intraclass correlation coefficient (two-way random effects model, ICC) between repeated measurements of the POM. Inter-rater and Intra-rater reliability was calculated using inter class correlations (ICC) consistency. Values > 0.7 are considered to be within acceptable range.

2.4.1.3. Measurement error: absolute measures. Measurement error is concerned with how close the scores on repeated measures are (Streiner & Norman, 2003). Agreement was determined by calculating the standard error of measurement (SEM) for consistency: $SD \times (1 - \sqrt{ICC_{consistency}})$ (Terwee et al., 2007). Thereafter the smallest detectable change (SDC) was calculated: $1.96 \times \sqrt{2} \times SEM$. The SDC needs to be smaller than minimal important change (MIC) to support agreement. The fraction of SDC and range of POM scores (minimal and maximum total scores) were compared relatively to the MIC.

2.4.2. Validity

2.4.2.1. Content validity. Content validity is assessed against the clarity of description of the measurement aim, the target population, the concepts that are being measured, the item selection, target population and investigators who were involved in item selection (Guyatt, Feeny, & Patrick, 1993).

2.4.2.2. Construct validity. Construct validity is defined as the extent to which scores on a particular instrument relate to other measures in a manner that is consistent with theoretical derived hypotheses concerning the concepts that are being measured (Terwee et al., 2007). To assess construct validity of the POM, the following two hypotheses were tested:

Hypothesis 1. There will be positive associations between the twelve overlapping POM and PP items.

Hypothesis 2. Given that children with ADHD are known to have pragmatic difficulties (Bignell & Cain, 2007), the participants with ADHD will have significantly lower mean measure scores on the POM than typically developing age, sex and ethnic matched peers.

2.4.2.3. Criterion validity. Criterion validity refers to the extent to which scores on a particular instrument are an adequate reflexion of a reference tool (Mokkink et al., 2010a, 2010b). Criterion validity was determined by comparing items from the PP and the POM that measured the same concept. To attain interval-level scores for each participant, PP and POM raw scores were subjected to Rasch analysis using the Winsteps programme (version 3.70.0.2; Linacre, 2007) respectively. This produced an overall PP and POM measure score for each participant. Criterion validity was assessed by: (a) calculating the correlations between the overall PP and POM measure scores and (b) calculating the correlations between overlapping PP and POM items (i.e., measuring the same construct).

2.4.3. Responsiveness

Responsiveness refers to the ability of an assessment to detect clinically important changes over time (Guyatt et al., 1993). The PP was used as the reference tool to measure responsiveness using PP and POM scores of both children with ADHD (with associated pragmatic language difficulties) as well as the typically developing children. The diagnostic categories of children were determined using 1.5 SD below the mean of PP total score as the diagnostic cut-off point (Beitchman et al., 2001). Subsequently, a discriminant analysis (DA) was performed to calculate sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood ratio and negative likelihood ratio. Receiver Operating Characteristic (ROC) curves analysis was used to further investigate the diagnostic test performance of the POM. The greater the area under the curve, the better the diagnostic test, since there will be smaller decreases in sensitivity as specificity increases and vice versa.

2.4.4. Interpretability

Interpretability is defined as the degree to which one can assign qualitative meaning to quantitative scores (Lohr et al., 1996) or change in scores (Mokkink et al., 2010a, 2010b). Interpretability can be investigated by: (a) comparing the means and SD of scores of relevant subgroups of participants who are expected to differ in scores (e.g., groups with different clinical diagnoses), and (b) comparing means and SD of scores of participants before and after treatment. In this study

interpretability was investigated by: (a) comparing the means and standard deviations of the overall POM measure scores of children with ADHD, with their typically developing playmates, and with age-, sex-, and ethnic-matched typically developing children in the control group from *Study 1*, and (b) comparing the pre-post intervention POM means and SD scores of 9 children with ADHD (from *Study 2*) who had associated pragmatic language difficulties and received a 7-week play-based intervention.

The POM was also investigated for floor or ceiling effects. Floor or ceiling effects are considered to be present if more than 15% of respondents achieved the lowest or highest possible score, respectively. The presence of floor and ceiling effects are indicative that extreme items are missing in the lower or upper end of the scale, suggesting limited content validity (Terwee et al., 2007).

3. Results

3.1. Reliability

3.1.1. Internal consistency

There were strong correlations between items within all the POM elements. The only item that had relative weak correlations was discourse interruption (Spearman $r_s = .351$; $p < .01$ to $r_s = .514$; $p < .01$). This suggests that discourse interruption may be a separate factor. Table 3 provides a summary of the range of correlational values for the respective POM dimensions.

An exploratory PC factor analysis was performed using all POM items. The factor analysis revealed two factors. The first factor comprised 27 items which explained 74.8% of the variance and a second factor, the discourse interruption item, which explained 6.7% of the variance. After removing the discourse interruption item, the results of the factor analysis suggested that the POM is a unidimensional construct, explaining 81.5% of the variance. A subsequent confirmatory Maximum Likelihood (ML) factor analysis was performed specifying one factor. This analysis confirmed that the POM was a unidimensional construct explaining 73.7% of the variance. Cronbach's α was calculated to be .987. Collectively, the findings of the factor analysis and the high Cronbach's α value strongly suggest that the POM is a unidimensional construct. Therefore a decision was made to remove the discourse interruption item.

3.1.2. Reliability

The inter-rater reliability was calculated for two independent raters using $ICC_{consistency}$ for the total POM scores of 42 participants: $r = .887$. The intra-rater reliability was calculated for the main rater using $ICC_{consistency}$ for the total POM scores of 40 participants: $r = .999$.

3.1.3. Measurement error: absolute measures

Standard Error of Measurement (SEM) of consistency was calculated: $SD \times (1 - \sqrt{ICC_{consistency}})$ (Terwee et al., 2007). The SEM of consistency for inter-rater reliability = .770. The SDC was calculated: $1.96 \times \sqrt{2} \times SEM$. The SDC of consistency for inter-rater reliability = 2.13. Therefore when comparing the SDC with the minimum and maximum POM scores (27 and 108, respectively), the SDC was considered too small to represent any clinically important change on a 4-point rating scale (minimum 2.13/27 = 0.079; maximum 2.13/108 = 0.020). The POM thus met the psychometric reproducibility property (SDC < MIC).

Using the same formulae used for inter-rater reliability, the SEM for consistency for intra-rater reliability was 0.042. The SDC for consistency for intra-rater reliability was 0.117. As with inter-rater reliability the SDC is considered to represent no clinical important change (minimum 0.117/27 = 0.004; maximum – 0.117/108 = 0.001; thus SDC < MIC). Therefore, given the small SDC for both inter- and intra-rater reliability, the SDC for consistency for the POM is accepted to be smaller than the MIC.

3.2. Validity

3.2.1. Content validity

A clear description was provided in the method section of the aim, the target population, the concepts that were being measured, the item selection, and investigators who were involved in item selection.

Table 3
Range of POM elements correlations.

POM elements	Range of correlations (r)[*]	
	Minimum	Maximum
Introduction and responsiveness	.782	.910
Non-verbal communication	.804	.874
Social-emotional attunement	.723	.907
Executive function	.764	.910
Negotiation	.624	.888

[*] $p < .001$ for all correlations.

Table 4
Correlations of overlapping PP and POM items.

PP item	POM item	Correlation[*]
Gestures	Gestures	.686
Body position	Body language	.639
Facial	Facial expressions	.702
Physical proximity	Distance	.671
Select	Select and introduce	.642
Introduce	Select and introduce	.770
Maintain	Maintain and change	.733
Change	Maintain and change	.663
Contingency	Contingency	.674
Initiate	Initiate	.788
Respond	Respond	.550
Repair and revise	Repair and revise	.700

[*] $p < .001$ for all correlations.

3.2.2. Construct validity

To assess construct validity of the POM, the following two hypotheses were tested:

Hypothesis 1. The hypothesis that there will be positive associations between the twelve (12) overlapping POM and the PP items was supported. All twelve items (Table 4) demonstrated a strong positive association ranging between $r = .550$ and $r = .770$.

Hypothesis 2. The hypothesis that children with ADHD from *Study 1* overall will have significantly lower mean measure scores on the POM than typically developing age, sex and ethnic matched peers was supported: independent sample t-test: Mean$_{ADHD}$ = 30.06 (SD = 37.67); Mean$_{Control}$ = 39.67 (SD = 34.31); $t = 2.03$; $df = 232$; $p = .043$.

3.2.3. Criterion validity

In calculating the criterion validity, a strong correlation between the *overall* PP and POM item measure scores was found (Pearson $r = .953$; $p = .005$), suggesting good criterion validity. Furthermore, strong associations between similar items from both measures were found. Table 4 summarises the correlations between overlapping PP and POM items. Collectively the results demonstrate that the PP and the POM defined the same target construct.

3.3. Responsiveness

The PP was used as the reference tool against which to calculate responsiveness. To determine if children had significant pragmatic language difficulties, the PP mean measure scores of the typically developing children ($M = 31.92$; $n = 126$) and 1.5 SD (23.9) below the mean measure score were used to determine the diagnostic cut-off point (8.02). Children with a mean measure score below 8.02 were then coded to have pragmatic language difficulties (PLD) and those above the cut-off where deemed not to have pragmatic language difficulties (noPLD). A discriminant analysis (DA) was then conducted using the POM measure scores of all children from *Study 1* ($N = 342$), against the diagnostic categories of PLD and noPLD using the diagnostic codes from the PP. The findings from the DA are summarised in Table 5.

The diagnostic test performance of the POM can be summarised as follows: sensitivity = 79.7%; specificity = 89.6%; positive predictive value = 67.8%; negative predictive value = 94.1%; likelihood ratio positive (LR+) = sensitivity/$(1 - \text{specificity})$ = 7.66; likelihood ratio negative (LK−) = $((1 - \text{sensitivity})/\text{specificity})$ = 0.23. ROC curves analysis was used to refine the analysis. The ROC analysis revealed that the POM covers .939 of the area under the curve ($p = <.001$; lower bound = .913; upper bound = .965 at 95% CI).

3.4. Interpretability

To investigate interpretability, the means and SD of the overall POM measures scores of children with ADHD, their typically developing playmates and typically developing children in the control group were compared (*Study 1*). As expected, the mean measure scores of children with ADHD were the lowest, followed by their playmates and typically developing

Table 5
Discriminant analysis of children with and without pragmatic language difficulties.

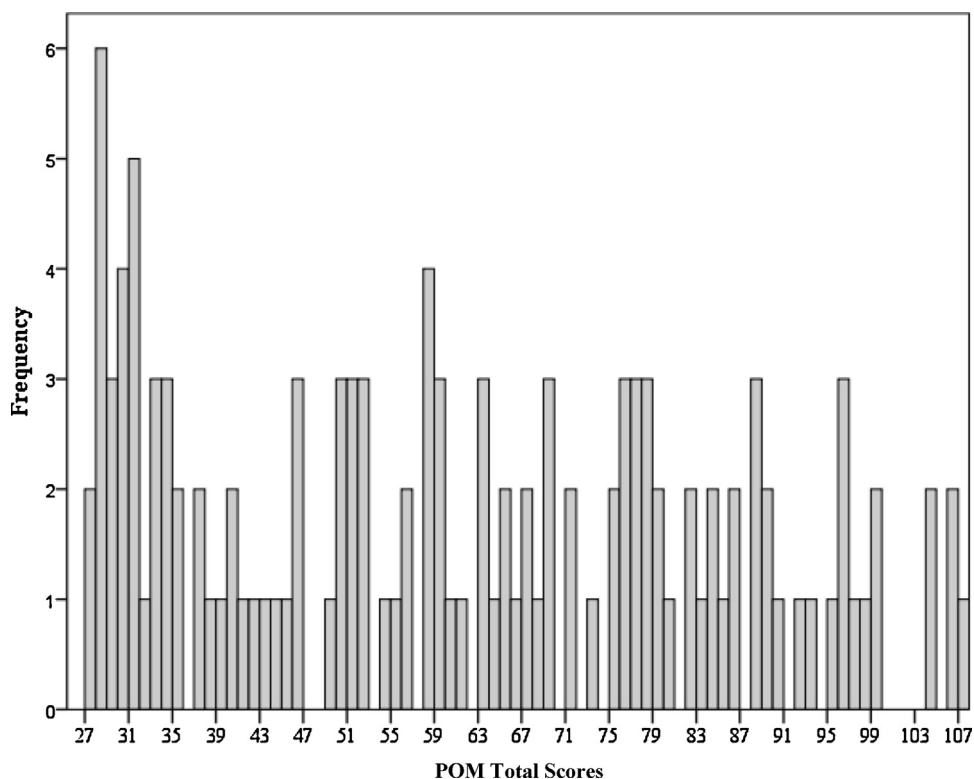|	PLD (n)	noPLD (n)
Test outcome positive	59	28
Test outcome negative	15	240

Fig. 1. Floor and ceiling effects of the POM total item scores.

children in the control group had the highest mean measure scores: POM $mean_{ADHD} = 36.55$, SD 35.36, $n = 108$; $mean_{Playmate} = 44.42$, SD = 32.13, $n = 108$; $mean_{Control} = 46.17$, SD = 29.92, $n = 126$. Furthermore, clinically important changes over time were investigated for nine children with ADHD who were subjected to a pilot 7-week parent delivered play-based intervention aimed at improving children's pragmatic language (*Study 2*) (Cordier et al., in preparation). Pre-post POM measure scores demonstrated that the children's overall pragmatic language scores improved significantly following the intervention using a paired sample *t*-test: pre mean = −11.57 (SD = 38.88); post mean = 44.99 (SD = 32.73); $t = 4.87$; $df = 8$; $p = .001$.

As there are 27 items with a minimum possible score of 1, and a maximum possible score of 4, the lowest possible sum of scores is 27 and the highest possible sum of scores is 108. Only 1.6% ($n = 2$) achieved the lowest possible score and 0.8% ($n = 1$) achieved the highest possible score. Given that there were 126 respondents, the findings demonstrate that less than 15% of respondents achieved the lowest or highest possible scores thereby supporting the absence of floor and ceiling effects. Fig. 1 shows the spread of total item scores.

4. Discussion

In this study, psychometric characteristics were determined for the POM, a new pragmatic assessment for school-aged children within the context of peer–peer play transactions. The COSMIN taxonomy of psychometric properties and definitions for health-related outcomes (Mokkink et al., 2010a, 2010b) was used as a framework against which the psychometric properties were assessed. The Pragmatic Protocol (PP) was considered the reference tool.

Internal-consistency for the POM was good. The findings of the factor analyses and the high Cronbach's α strongly support that the POM measures a unidimensional construct: pragmatic language of school-aged children within the context of peer–peer play. This suggests all 27 items relate to the same construct–pragmatic language. The $ICC_{consistency}$ inter-rater and intra rater reliability were within acceptable range. The smallest detectable change for POM scores was very small, indicating an acceptable level of measurement error.

The domain of interest was comprehensively sampled by the items of the POM, thus indicating good content validity. Good construct validity was determined as the POM data were in accordance with theoretically derived hypotheses that had been formulated prior to examining the psychometric properties of the POM.

Strong correlations were found between the overall PP and the POM item measures suggesting good criterion validity. The criterion validity between similar items from both measures ranged between moderate and good. The PP scores of typically developing children were used as the reference tool to determine the test performance of the POM for diagnosing

children with pragmatic language difficulties. High sensitivity and specificity scores, together with ROC curves analysis indicated that the POM is responsive to clinical changes over time.

Pre- and post-intervention POM measure scores in a group of children with ADHD (*Study 2*) demonstrated clinical and statistical significant improvement over time. This suggests that the POM is capable of distinguishing between children with and without pragmatic difficulties in a sample of children with ADHD and typically developing children within the context of peer–peer interactions. Furthermore the POM can track clinical changes over time, thus supporting the interpretability of the POM. No floor or ceiling effects were found for the POM scores.

4.1. Limitations and implications for future research

Sample numbers were relative small for evaluating responsiveness and interpretability due to the small number of children in the intervention group (*Study 2*). Use of the POM in larger more robust intervention studies will be important for demonstrating responsiveness with respect to therapy outcomes. The psychometric properties of the POM need to be assessed in other clinical groups likely to have pragmatic language difficulties, including autism spectrum disorders, and a range of other communication, behavioural, cognitive and learning disorders. Future research could also look at the relationship between the POM and instruments that assess peer socialisation in naturalistic contexts such as the Manchester Inventory for Playground Observation (MIPO: Gibson, Hussain, Holsgrove, Adams, & Green, 2011).

5. Conclusions

Considering all analyses, it can be concluded that the POM demonstrated good psychometric characteristics. The psychometric properties of the POM suggest that it is a reliable and valid measure of pragmatic language skills of children with ADHD between the age of 5 and 11 years within the context of peer–peer play. Furthermore, the POM demonstrated that it has clinical utility in identifying children with pragmatic language difficulty. This is important as the POM is designed to be a clinical tool that aids in reliably identifying children, aged between 5 and 11 years as having pragmatic language difficulties during peer–peer play, and in tracking clinical change over time.

Whilst the PP was important in serving as the reference tool, its binary scoring system proved to be clumsy in detecting the subtlety associated with pragmatic language abilities and was therefore considered to have more limited clinical utility. Conversely the 4-point scoring system of the POM, together with its application of a contemporary definition of pragmatics, demonstrates great promise.

There is a clear need for the development of psychometrically sound pragmatic language assessments (Gerber et al., 2012). In this paper, we have argued for the need to develop an observational assessment of peer–peer pragmatic language during naturalistic play interactions. The use of the social context of peer–peer interaction and play for school-aged children is both clinically and theoretically motivated (Cordier et al., 2013). Establishing the POM's psychometric properties is the first step towards contributing to this neglected but much-needed area of research.

References

Adams, C. (2002). Practitioner review: The assessment of language pragmatics. *The Journal of Child Psychology and Psychiatry, 43*(8), 973–987.
Adams, C., Baxendale, J., Llyod, J., & Aldredge, C. (2005). Pragmatic language impairment: Case studies of social and pragmatic therapy. *Child Language Teaching and Therapy, 21*, 227–250.
Adams, C., Lockton, E., Freed, J., Gaile, J., Earl, G., McBean, K., et al. (2012). The Social Communication Intervention Project: A randomized controlled trial of the effectiveness of speech and language therapy for school-age children who have pragmatic and social communication problems with or without autism spectrum disorder. *International Journal of Language and Communication Disorders, 47*(3), 233–244.
Adams, C., Lockton, E., Gaile, J., & Freed, J. (2011). TOPICCAL applications: Assessment of children's conversation skills. *Speech and Language Therapy in Practice, Spring*, 7–9.
American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders (revised)* (4th ed.). Washington, DC: American Psychiatric Association.
Beitchman, J. H., Wilson, B., Johnson, C. J., Atkinson, L., Young, E. M., Adlaf, E., et al. (2001). Fourteen-year follow-up of speech/language-impaired and control children: Psychiatric outcome. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*(1), 75–82.
Bignell, S., & Cain, K. (2007). Pragmatic aspects of communication and language comprehension in groups of children differentiated by teacher ratings of inattention and hyperactivity. *British Journal of Developmental Psychology, 25*, 499–512.
Bishop, D. (2003). *The children's communication checklist-2*. London: Psychological Corporation.
Brinton, B., & Fujiki, M. (2006). Social intervention for children with language impairment: Factors affecting efficacy. *Communication Disorders Quarterly, 28*(1), 39–41.
Conners, C. (2004). Validation of ADHD rating scales: Dr. Conners replies. *Journal of the American Academy of Child and Adolescent Psychiatry, 43*(10), 1190–1191.
Conners, C. K., Sitarenios, G., Parker, J. D. A., & Epstein, J. N. (1998). The revised Conners' Parent Rating Scale (CPRS-R): Factor structure, reliability, and criterion validity. *Journal of Abnormal Child Psychology, 26*, 257–268.
Cordier, R., & Bundy, A. C. (2009). Children and playfulness. In K. Stagnitti & R. Cooper (Eds.), *Play as therapy: Assessment and therapeutic interventions.* London, United Kingdom: Jessica Kingsley Publishers.
Cordier, R., Bundy, A., Hocking, C., & Einfeld, S. (2010a). Playing with a child with ADHD: A focus on the playmates. *Scandinavian Journal of Occupational Therapy, 17*(3), 191–199.
Cordier, R., Bundy, A., Hocking, C., & Einfeld, S. (2010b). Empathy in the play of children with ADHD. *OTJR: Occupation, Participation and Health, 30*(3), 122–132.
Cordier, R., Ling, L., Munro, N., & Wilkes-Gillan, S. (in preparation). Pragmatic language outcomes following a parent-delivered play-based intervention: Pilot findings.
Cordier, R., Munro, N., Wilkes, S., & Docking, K. (2013). The pragmatic language abilities of children with ADHD following a play-based intervention involving peer-to-peer interactions. *International Journal of Speech-Language Pathology, 5*(4), 429–440.

Feiring, C., & Lewis, M. (1989). The social networks of girls and boys from early through middle childhood. In D. Belle (Ed.), *Children's social networks and social supports* (pp. 119–150). New York: Wiley-Interscience.

Florey, L., & Greene, S. (2008). Play in middle childhood. In L. Parham & L. Fazio (Eds.), *Play in occupational therapy for children* (2nd ed., pp. 279–299). St. Louis, MO Mosby.

Fujiki, M., Brinton, B., & Clarke, D. (2002). Emotion regulation in children with specific language impairment. *Language, Speech, and Hearing Services in Schools, 33,* 102–111.

Gallagher, T. M. (2012). Social/pragmatic language intervention with young children. *Journal of the Speech-language Hearing Association of Taiwan, 29,* 45–55.

Gerber, S., Brice, A., Capone, N., Fujiki, M., & Timler, G. (2012). Language use in social interactions of school-aged children with language impairments: An evidence-based systematic review of treatment. *Language, Speech and Hearing Services in Schools, 43,* 235–249.

Gibson, J., Hussain, J., Holsgrove, S., Adams, C., & Green, J. (2011). Quantifying peer interactions for research and clinical use: The Manchester Inventory for Playground Observation. *Research in Developmental Disabilities, 32,* 2458–2466.

Guyatt, G. H., Feeny, D. H., & Patrick, D. L. (1993). Measuring health related quality of life. *Annals of Internal Medicine, 118,* 622–629.

Hart, K., Fujiki, M., Brinton, B., & Hart, C. (2004). The relationship between social behavior and severity of language impairment. *Journal of Speech, Language and Hearing Research, 47,* 647–662.

Jerome, A., Fujiki, M., Brinton, B., & James, S. (2002). Self-esteem in children with specific language impairment. *Journal of Speech, Language and Hearing Research, 45,* 700–714.

Leonard, M. A., Milich, R., & Lorch, E. P. (2011). The role of pragmatic language use in mediating the relation between hyperactivity and inattention and social skills problems. *Journal of Speech, Language and Hearing Research, 54,* 567–579.

Linacre, J. M. (2007). *Winsteps Rasch measurement: Computer software manual (version 3.70.0.2)* Retrieved from http://www.winsteps.com/index.htm.

Lohr, K. N., Aaronson, N. K., Alonso, J., Burnam, M. A., Patrick, D. L., Perrin, E. B., et al. (1996). Evaluating quality of life and health status instruments: Development of scientific review criteria. *Clinical Therapeutics, 18*(5), 979–992.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010a). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63,* 737–745.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010b). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research, 19,* 539–549.

Prutting, C., & Kirchner, D. (1987). A clinical appraisal of the pragmatic aspects of language. *Journal of Speech and Hearing Disorders, 52,* 105–119.

Russell, R., & Grizzle, K. (2008). Assessing child and adolescent pragmatic language competencies: Toward evidence-based assessments. *Clinical Child and Family Psychology Review, 11,* 59–73.

Semel, E., Wiig, E., & Secord, W. (2003). *Clinical Evaluation of Language Fundamentals —Fourth Edition, Pragmatics Profile.* San Antonio, TX: The Psychological Corporation.

Staikova, E., Gomes, H., Tartter, V., McCabe, A., & Halperin, J. M. (2013). Pragmatic deficits and social impairment in children with ADHD. *Journal of Child Psychology and Psychiatry, 54*(12), 1275–1283.

Statistics New Zealand. (2007). *National population estimates June 2007 quarter* Retrieved from http://www.stats.govt.nz/products-and-services/hot-off-the-press/national-population-estimates/national-population-estimates-jun07-qtr-hotp.htm.

Streiner, D. L. (2003). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment, 80,* 217e22.

Streiner, D. L., & Norman, G. R. (2003). *Health measurement scales. A practical guide to their development and use.* New York: Oxford University Press.

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60,* 34–42 http://dx.doi.org/10.1016/j.jclinepi.2006.03.012

Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W. J. G., Bouter, L. M., & de Vet, H. C. W. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. *Quality of Life Research, 21,* 651–657.

Wiig, E., Larson, V., & Olson, J. (2004). *S-MAPs: Curriculum-based assessment and intervention.* WI, USA: Thinking Publications.