Chapter 19
Game-Based Spoken Interaction Assessment in Special Need Children

Jos Keuning, Sanneke Schouwstra, Femke Scheltinga and Marleen van der Lubbe

Abstract The purpose of the present study was to explore whether it is possible to collect high-quality data about children's spoken interaction skills using the *Fischerspiel* board game as an entertaining, non-threatening means to evoke conversations between children in special elementary education. The game was administered to a total of 681 eleven- and twelve-year-old children with varying educational needs. The quality of the conversations between the children was evaluated with a specially designed observation form. The observation forms were filled in by trained test leaders and four independent expert raters. Video recordings showed that almost all children were willing to participate in the game, even the children who usually barely speak in class. Moreover, the game provided more than sufficient data to assess different dimensions of spoken interaction skills. Analyses further showed that the observation form functioned well and provided reliable scores. A group effect was nevertheless present and two test leaders deviated largely from the expert raters. These test leaders may have been insufficiently equipped to cope with the task. Application of Automatic Speech Recognition (ASR) technology in a (computer-based) spoken interaction assessment might ease the task and increase rating quality.

19.1 Introduction

Oral language performance is a good predictor of various school outcomes such as mathematics (Chow and Jacobs 2016; Fuchs et al. 2005) and reading and writing (Dickinson et al. 2010; Hart and Risley 2003; Kent et al. 2014; Nation and Snowling 2004; Dougherty 2014). Deficits in oral language skills may also underlie difficulties

J. Keuning (✉) · S. Schouwstra
Cito, Arnhem, The Netherlands
e-mail: jos.keuning@cito.nl

F. Scheltinga
University of Amsterdam, Amsterdam, The Netherlands

M. van der Lubbe
Inspectie van het Onderwijs, Utrecht, The Netherlands

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_19

in text comprehension and writing (Shanahan 2006; Naucler and Magnusson 2002). At school, language is necessary to understand instruction, to communicate about content and to demonstrate understanding. Moreover, language ability is related to social behavior and poor language skills affect social behavior negatively (Chow and Wehby 2018). Within the educational context, oral language is generally referred to as listening and speaking. Listening is as a receptive skill: it is the process of recognizing words and understanding messages expressed by others. Speaking is a productive skill: it is the process of transmitting ideas and information orally in a variety of situations. Both skills are used alternately in spoken interaction. Conversation partners must not only be able to convey and understand the content, but in addition, they must also be able to manage, produce and interpret verbal and non-verbal communication. While there is much experience in monitoring the listening and speaking skills of children, it is less clear how to monitor the children's spoken interaction skills. The present study focuses on an even more unexplored area: spoken interaction assessment in special need children.

19.1.1 Measuring Spoken Interaction Skills

Spoken interaction includes three content domains, namely Language form, Language content and Language usage (Bloom and Lahey 1978; Lahey 1988). Language form refers to the grammar of language, that is, to the phonological, morphological and syntactic rules that determine how sounds, words and sentences are formed, respectively. Language content refers to the ability to make up an own (non-)fictional story with a clear line of thought. Finally, in Language usage it is about the understanding and use of communicative functions and conversation rules such as expressing and interpreting emotion, making conversation, maintaining a topic of conversation, taking turns or asking others for information (McLaughlin 1998). When measuring spoken interaction skills it is important that all three content domains are covered. In addition, O'Malley and Pierce (1996) state that a spoken interaction assessment should: (a) test a child's competence as authentically as possible, (b) include an evidence model which shows how responses are scored, analyzed and interpreted, (c) take a limited amount of time to administer. In practice, it is difficult to meet all aforementioned criteria at the same time. Moreover, an extra complicating factor in assessing spoken interaction is the influence of the context and conversation partner. An interplay does exist between individual language ability and contextual factors (Chapelle 1998). Furthermore, the course of a conversation is determined in interaction and co-construction (Kramsch 1986).

Taylor and Galazci (2011) suggested to control context and to use detailed scripts to challenge children at their own level in a meaningful and standardized manner. These suggestions were implemented in a recent large-scale assessment in the Netherlands by having three children jointly conduct a conversation assignment (Van Langen et al. 2017). The assignment was conducted in the context of a national charity action: the children had to make up a goal and activity with which they could

Table 19.1 Spoken interaction assignments for one-to-one settings

1. Short situation sketch with an interaction between child and test leader + Reasonably standardized administration and scoring + Structured and therefore in line with the child's needs + Alignment on conversation partners and Taking turns are slightly covered − Children can suffice with short responses − It requires a lot from the test leader (or teacher) to respond in a standardized manner; outcomes are a bit uncertain
2. Role-play + Alignment on conversation partners and Taking turns are covered − According to teachers not suited for many special need children − Administration and scoring not standardized
3. Cartoon storytelling + Reasonably standardized administration and scoring + Elicits somewhat longer stories in most children − Alignment on conversation partners and Taking turns are not well covered − Appeals to the child's imagination and creativity − The performances in mainstream and special need education are often similar; this is not what is expected, validity might be an issue
4. Referential communication task + Reasonably standardized administration and scoring + Alignment on conversation partners and Taking turns are slightly covered − Appeals to specific concepts or words (e.g., colors, shapes, front/back/left/right etc.) − The child only gives an instruction, storytelling is not really part of the assignment

raise money. In order to finance the preparation of their idea they were awarded a (fictitious) amount of 50 euro. The children had to agree on the project planning and the corresponding financial framework. The topics to be discussed were shown on a paper for the children as a guide during the assignment. In addition, each child received his own conversation guide as input for the conversation. The conversation guide contained information that was specifically intended for that child. For example, each child was expected to have his or her own contribution at various points in the conversation. The conversation guides encouraged each child to actively contribute to the conversation at minimally two moments. Although not every child received information on all discussion topics, they always did have the opportunity to contribute. The child could, for example, give an opinion, give a reaction to the opinion of others, join this opinion or make a compromise. In this manner, the conversation could proceed as naturally as possible.

The charity assignment functioned well when administered to mainstream eleven- and twelve-year-old children (see Van Langen et al. 2017), but it is unlikely that the assignment is also appropriate for special need children. According to special education teachers, the context is, for instance, too abstract for children with lower cognitive abilities. Many other spoken interaction assignments are possible, all with specific advantages and disadvantages. Tables 19.1, 19.2 and 19.3 show which assignments can be administered in a one-to-one setting, a small-group setting or a classroom setting, respectively.

Table 19.2 Spoken interaction assignments for small-group settings

1. Group storytelling (e.g. about the weekend) + Alignment on conversation partners and Taking turns are covered dependent on the role of the child in the group − Some children may contribute a lot while others do not; the group affects the chance an individual child gets to show what he or she can do − Administration and scoring are not standardized
2. Group assignment (e.g. development of a project plan) + Alignment on conversation partners and Taking turns are covered dependent on the role of the child in the group − Some children may contribute a lot while others do not; the group affects the chance an individual child gets to show what he or she can do − Administration and scoring are not standardized − Not well suited for special need children

Table 19.3 Spoken interaction assignments for classroom settings

1. Lecture or book discussion + The child can show extensively what he or she can do − Alignment on conversation partners and Taking turns are not well covered − Administration not standardized
2. Observation in a natural situation such as a circle discussion + A familiar situation for the children − Administration and scoring are not standardized − Some children may contribute a lot while others do not − Not all children may feel safe in a circle discussion

19.1.2 Assessment in Special Elementary Education

At this moment, it is unclear which assignment should be preferred in which situation. The assignment must, however, account for the specific characteristics of the children in special elementary education. Special elementary education is meant for children who need orthopedagogical and ortho-didactical support. In the Netherlands, almost all special need children experience multiple problems in relation to learning, behavior or development. The children show problematic internalizing or externalizing behavior, for example, or have a communication problem or intellectual disability (see, for example, Ledoux et al. 2012). It is the accumulation of problems that makes it difficult, and sometimes impossible, for these children to follow education in a regular setting. A few considerations apply when developing an assignment or test for special need children. First, the children's special needs are generally better taken into account in practical assignments than in paper-and-pencil multiple-choice tests. Second, due to the limited attention span of many special education children, assignments should be short and varied. Third, assessments should engage the children and give them a feeling of success as many already encountered multiple failure experiences during their school career. Finally, the children in special education show great diversity and some children require (additional) adjustments to assessment practices in order to demonstrate what they know and can do. A protocol with an overview of allowable adjustments is required to cater for the characteristics of the children being assessed.

An assessment in special elementary education also requires specific choices with regard to context, layout and use of language (see, for example, Cito 2010). Contexts must be meaningful and connect to the children's experiences, for instance, and potentially 'provocative' contexts such as a football match must be avoided at all times. Images must have a high contrast and be available to the children in black-and-white or enlarged format. Moreover, the images must be 'real' and support the assignment; talking animals or purple frogs are not appropriate, for example. Finally, language should be as concrete as possible. Negative sentences, imagery and emotionally charged sentences such as 'stop doing that' should be avoided, just as I-sentences, long compound sentences and complex cause-and-effect relationships. Although such guidelines also apply to children in regular elementary education to some degree, they are particularly important in special elementary education. Especially the children with special educational needs will perform best in familiar situations without pressure. In general, the children will naturally apply many of their listening and speaking skills in free circumstances. Data about the children's skills can then be collected best by observation; it does not bother the children with an assessment and they can show their skills in a familiar setting without 'explicitly knowing' that their skills are being monitored and documented.

19.1.2.1 The Present Study

In recent literature, game-based assessment has been developed as a more fun and accessible way for children to assess their knowledge and skills. Games are well suited for assessment purposes as they naturally present children with a stream of choices during gameplay. All these choices can be recorded and it is also possible to record how the children arrived at their choice (Stieger and Reips 2010). This allows game-based assessments to capture information that often cannot be captured by traditional paper-and-pencil assessments (Shute and Ventura 2013; Landers 2014). Moreover, it is relatively easy to create authentic and familiar situations as children play games every day. However, especially in special elementary education it should not be a battle against each other; solving the game together should be the objective. Against this background, a game for the assessment of spoken interaction skills was developed. The game was based on the *Fischerspiel*. This is essentially a board game which will be described in more detail below. An observation form was further developed in order to assess the children's spoken interaction skills during gameplay.

Observation as a measurement strategy has, despite numerous advantages, certain unique limitations (Michaels 1983), such as imposed limitations on the types of behavior observed, problems with the category systems, observer bias and interferences. In this study it was examined whether such limitations occurred in the games-based assessment for spoken interaction. Different aspects of the game, the observation form and the test leader were evaluated. The first objective was to evaluate whether particular game characteristics were a source of invalidity. Messick (1995) distinguished two sources of invalidity: underrepresentation and irrelevant variance. When an assessment fails to include important aspects of the skill the assessment suffers from underrepresentation. When an assessment contains excess variance associated with other aspects than the skill of interest the assessment is hampered by irrelevant variance. Both sources of invalidity were studied: the variety in the children's conversations was mapped, and in addition, it was examined whether the group and turn-taking affected the spoken interaction skills that individual children showed. The quality of the observation form was evaluated next. It was examined whether the assessment covered the relevant aspects of spoken interaction, and moreover, scale dimensionality was examined via exploratory factor analysis. Finally, the third important aspect of the assessment was considered: the test leader. The quality and reliability of the test leaders' evaluations were mapped by comparing the test leader ratings to an expert rating. The overall quality of the ratings was assessed and it was attempted to identify extreme or deviating ratings.

19.2 Method

19.2.1 Participants

A total of 681 eleven- and twelve-year-old children from 33 different special education schools in the Netherlands participated in the study. A two-fold stratification procedure was used to select the schools. Region was used as explicit stratification criterion: all Dutch special education schools were classified by region (North, East, South and West) and then a separate sample was drawn for each of the regions, so that the relative share of each region in the sample was representative of the relative share in the population of Dutch special education schools. School size was used as implicit stratification criterion: within each region the schools were organized from small to large and then, after generating a random start point, every *k*th school on the list was selected, so that both smaller and larger schools were included in the sample. No exclusion criteria were used for drawing the school sample. Within each school all children in the final grade (eleven- and twelve-year-olds) were expected to participate. Children with specific language impairment, hearing problems, selective mutism or aphasia were excluded from the study and also the children who lived in the Netherlands for less than 2 years were not eligible to participate. The sample consisted of 423 boys (62%) and 258 girls (38%) at varying educational levels. It was expected that after elementary school about 7% of the children would move on to General Secondary Education or higher. The other children would expected to move on to either Preparatory Vocational Education (49%) or Special Secondary Education (44%). These percentages are in line with the Dutch special education school population. More boys than girls attend Dutch special education and only a very small percentage moves on to the higher levels of Dutch secondary education.

19.2.2 Materials

The children's spoken interaction skills were assessed with an existing board game from Germany; the *Fischerspiel*. The game is played on a board with an island with several harbors and a sea. Each player has his or her own harbor and a colored boat to transport fish from the sea to the harbor. The aim of the game is to work together to bring all fish to the harbor before the wind reaches strength 12. When a player gets a turn, he throws a special die and consults his fellow players to determine who can best use the thrown number to get a fish and bring it to one of the harbors on the island. Players win together if all the fish are on the island. There is also a wind symbol on the die, however, and rolling the wind symbol increases the strength of the wind by 1. When the wind reaches strength 12, all boats sink and the game is lost. The quality of the conversations between players was evaluated with a specially designed observation form. The form included seventeen performance aspects. Each performance aspect was presented with three indicators: poor basic proficiency (0); fair proficiency (1) and good basic proficiency (2). Below the seventeen indicators of a good basic proficiency level are presented:

1. The child's conversations with the group are meaningful and relevant.
2. The child regularly takes the initiative to start, continue or stop a conversation.
3. The child usually takes the floor in an appropriate way.
4. The child integrates contributions from the group into his own contribution when relevant.
5. The child takes the initiative to achieve a joint communication goal by involving the group in the conversation.
6. The child makes his way of thinking understandable.
7. The child consistently uses language that fits the situation.
8. The non-verbal behavior of the child strengthens his verbal message.
9. The child shows adequate active listening behavior.
10. The child consistently gives appropriate verbal and nonverbal responses.
11. The child's contribution shows sufficient variation in word use
12. The child's vocabulary is sufficient to hold a conversation.
13. The child speaks fairly fluently with only occasional hitch, false starts or reformulation.
14. The child's pronunciation, articulation and intonation make the child's speech intelligible, despite a possible accent.
15. The child conjugates verbs correctly.
16. The child uses (combinations with) nouns correctly.
17. The child generally constructs correct simple, complex and compound sentences.

The observation form was a reflection of the Dutch reference framework for spoken language (Meijerink 2009). At the basic level of spoken language proficiency (1F) it is, for instance, expected that the child recognizes conversation situations and can use appropriate routines to give instruction or exchange information. At the highest level (4F) it is expected that the child is able to participate in casual, formal, and extended conversations on practical and academic topics. Language levels 1F and 2F apply to (special) elementary education.

19.2.3 Procedure

Administration of the *Fischerspiel* board game took place in a small and quiet, relatively stimulus-free room. The game was played in groups of three to four children. The groups were assembled randomly, but if a combination of children was inconvenient according to the teacher, a small change in the composition of the group was allowed. A quarrel during the break could, for example, be a reason to place a child in another group. Each administration was supervised by a test leader. The test leader did not participate in the game but acted as coach. The test leader monitored the course of the game and ensured that all the children got an equal number of turns and felt safe. In addition to a coaching role, the test leader also fulfilled the role of assessor during the administration. The observation form was filled in for each child separately after three rounds of the game. Try-outs showed three playing rounds to be more than sufficient to get an idea of the children's spoken interaction skills. Moreover, the children generally could play the game independently after three rounds, giving the test leader time to fill in the forms. In order to ensure that the test leaders could conduct the assessment task as reliably as possible, the following four measures were taken:

1. Each performance indicator was elaborated with one or more examples.
2. The test leaders received an extensive training on the use of the assessment form.
3. Each administration was recorded on video, so that the test leader had the possibility to complete or check the assessment afterwards.
4. Questions about the assessment and dilemmas could be presented to other test leaders in a WhatsApp group.

The administration of the *Fischerspiel* board game took approximately 30 min, depending on the course of the game. To prevent potential group effects and effects of turn-taking order the following was done:

(a) Children were randomly assigned into groups.
(b) There was a starting round and each child had several turns; at a certain moment it is unlikely that the children still know who exactly started.
(c) The child who had the turn always had to take the initiative, but other players had the possibility to respond; there was no fixed order.

After completion of the administrations, a selection of children were re-assessed by a subject-area expert. The re-assessment was conducted in an incomplete design which was specifically developed to efficiently detect aberrant rating behavior. The design assumed that there were b, $b = 1,\ldots, B$, test leaders and m, $m = 1,\ldots, M$, expert assessors. From each test leader b a total of J children were selected (1 per

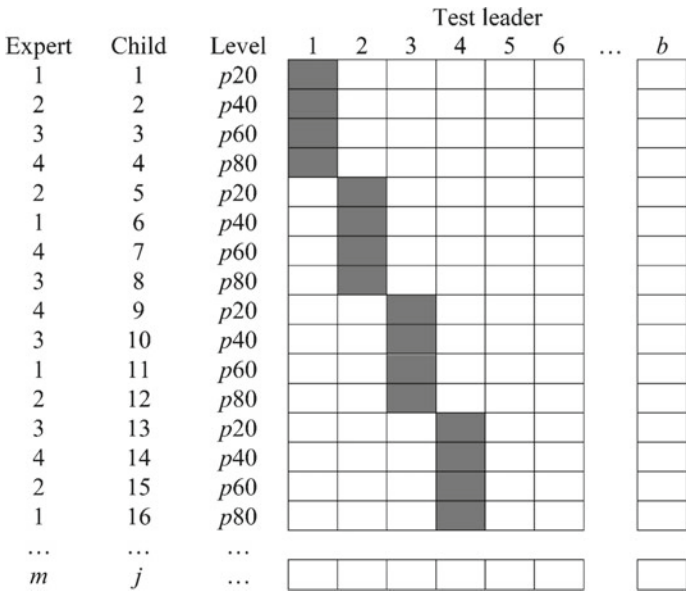


Fig. 19.1 Schematic representation of the design used to examine rater reliability

group) and each expert assessor m re-assessed $B \times J$ children. The children j were selected on the basis of their percentile rank in order to ensure that both low and high ability children were re-assessed. A total of 16 test leaders was involved in this study, and four different subject-area experts all re-assessed one child per test leader. This means that a total of 64 children (16 × 4) were re-assessed by one of the four subject-area experts. Figure 19.1 gives a schematic representation of the design. As soon as the re-assessments were conducted, difference scores were calculated for each performance indicator by subtracting the expert rating score form the test leader rating score. The difference scores were then the basic observation in the analysis.

19.2.4 Statistical Analyses

Analyses within the framework of Classical Test Theory were conducted to answer the first research question. First the distribution of total scores was examined and then for each of the seventeen performance aspects (items) the p-value and r_{it}-value was computed. The p-value was computed as the ratio between the mean score and the maximum achievable score. Values between 0.500 and 0.700 can be considered optimal (Crocker and Algina 1986; Feldt 1993), but lower (>0.100) and higher values (<0.900) might be acceptable dependent on item type and purpose of the test. The r_{it}-value is the correlation between the item scores and total scores. Values below 0.190 indicate that the item does not discriminate well, values between 0.200 and

0.290 indicate sufficient discrimination, and values of 0.300 and above indicate good discrimination (Ebel and Frisbie 1991). Although the Classical Test Theory analyses can easily be conducted, the manner of administration may distort results. Separate analyses were therefore conducted to examine whether group or turn order effects were present or not. Classical Test Theory analyses were conducted separately for the first, second, third and fourth group member, and by means of a three-level regression analysis the proportion of variance explained by group membership was estimated.

To answer the second research question, the matrix of polychoric correlations between the seventeen performance aspects was visually inspected by means of a correlogram. After inspection of the correlogram, an exploratory Principal Axis Factor Analysis with Varimax rotation was conducted. In order to choose the number of factors well-reasoned, we started with a parallel analysis as proposed by Horn (1965): a simulation-based method in which essentially a random dataset is generated with the same number of items and exactly the same score range. The eigenvalues of the items in this random simulated dataset are then compared with the eigenvalues of the items in the actual dataset. All factors with an eigenvalue larger than the random (simulated) eigenvalues were retained in the factor analysis. The viability of the factor solution was assessed in light of the Dutch reference framework for spoken language and the conceptual framework by Bloom and Lahey (1978) and Lahey (1988).

The third research question was answered by examining the reliability and quality of the rating scores. Reliability was estimated in terms of the Greatest Lower Bound and Guttman's Lambda2 (Sijtsma 2009; Ten Berge and Sočan 2004). Coefficients higher than 0.800 were considered to be good and coefficients below 0.700 to be insufficient. The differences between the expert rating scores and the test leader rating scores were used to evaluate the quality of the rating. The lack of agreement with the norm (i.e., the expert rating scores) was mapped for each of the test leaders by computing the Mean Absolute Error (MAE):

$$MAE_b = \frac{\sum_j \sum_i \left| s_{bji} - s_{mji} \right|}{N_j},$$

where s_{bji} and s_{mij} are the rating scores of test leader b and expert m, respectively, for child j on item i and N_j the number of ratings on child j by test leader b and expert m. The Median Absolute Deviation (MAD) was used as measure for detecting aberrant rating behavior. To optimally account for a possible asymmetric distribution of **MAE**, the median absolute deviation from the median was based on all points greater than or equal to the median: $MAD = Mdn(|\mathbf{Y} - Mdn(MAE)|)$, where $\mathbf{Y} = \{MAE_b \in \mathbf{MAE} : MAE_b \geq Mdn(\mathbf{MAE})\}$. Given this distance, a test leader b was marked as outlier if:

$$\frac{MAE_b - Mdn(\mathbf{MAE})}{MAD} > 2.5$$

Threshold value 2.5 was suggested by Leys et al. (2013) but other values are possible. The overall quality of the ratings was finally assessed by computing

Cohen's weighted kappa coefficient (κ) and Gower's similarity coefficient (G_{xy}). Cohen's kappa was interpreted as follows: κ < 0.200 poor; 0.200 < κ < 0.400 fair; 0.400 < κ < 0.600 moderate; 0.600 < κ < 0.800 good; κ ≥ 0.800 excellent. Gower's similarity coefficient was considered low if G_{xy} < 0.650, acceptable if 0.650 ≤ G_{xy} < 0.800, and high if G_{xy} > 0.800.

19.3 Results

The distribution of total scores is visually presented on the left-hand side in Fig. 19.2. As can be seen, the score distribution was slightly skewed to the left (−0.814) and had fatter tails than a normal distribution (3.155); the sample mean was 25.790 with a standard deviation of 6.260. On the average, children obtained about three quarters of the total number of points that could maximally be achieved (25.790 ÷ 34). Further analysis showed that the seventeen items functioned quite similarly. As can be seen from Table 19.4, the *p*-values varied from 0.596 to 0.916 and the r_{it}-values were all higher than 0.300. Although the size of the r_{it}-values is related to the manner in which the items were scored (i.e., a three-point Likert scale instead of dichotomous correct-incorrect scores), the r_{it}-values indicated the items to discriminate very well. Columns c1, c1, and c2 show the percentage of children with scores 0, 1 and 2 respectively. The number of children with a zero score was remarkably low for some items, especially for those related to fluency, comprehensibility or grammar mastery. Table 19.4 nevertheless shows that the items in the observation form were all appropriate for distinguishing children with weaker spoken interaction skills from children with better spoken interaction skills. There were no reasons to drop items from the observation form or to merge score categories.

The Classical Test Theory analyses were repeated for the first, second, third and fourth group member separately in order to examine whether turn order effects were

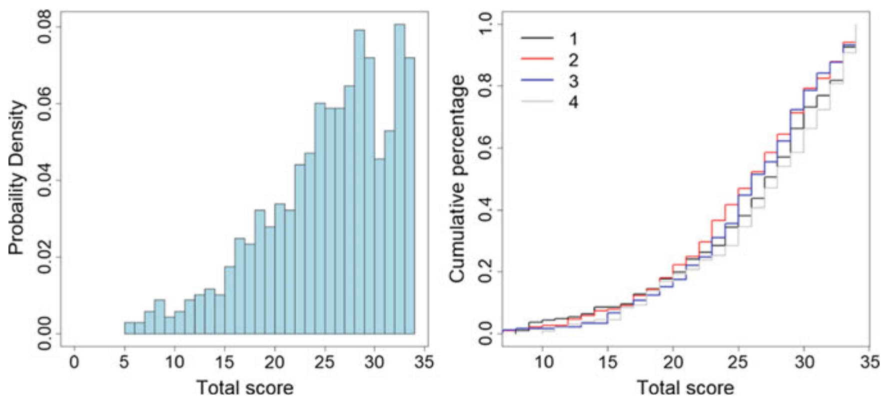


Fig. 19.2 Score distribution for the entire sample (left) and each group member (right)

Table 19.4 Results analyses at item level

Item	c0	c1	c2	p	r_{it}
1. The child's conversations with the group are meaningful and relevant	0.157	0.493	0.349	0.596	0.725
2. The child regularly takes the initiative to start, continue or stop a conversation	0.157	0.441	0.402	0.623	0.575
3. The child usually takes the floor in an appropriate way	0.098	0.276	0.626	0.764	0.539
4. The child integrates contributions from the group into his own contribution when relevant	0.072	0.266	0.662	0.795	0.538
5. The child takes the initiative to achieve a joint communication goal by involving the group in the conversation	0.164	0.446	0.389	0.612	0.638
6. The child makes his way of thinking understandable	0.116	0.382	0.502	0.693	0.741
7. The child consistently uses language that fits the situation	0.035	0.305	0.659	0.812	0.613
8. The non-verbal behavior of the child strengthens his verbal message	0.072	0.606	0.322	0.625	0.520
9. The child shows adequate active listening behavior	0.106	0.461	0.433	0.664	0.679
10. The child consistently gives appropriate verbal and nonverbal responses	0.110	0.285	0.605	0.747	0.574
11. The child's contribution shows sufficient variation in word use	0.070	0.256	0.674	0.802	0.729
12. The child's vocabulary is sufficient to hold a conversation	0.031	0.223	0.746	0.858	0.698
13. The child speaks fairly fluently with only occasional hitch, false starts or reformulation	0.040	0.270	0.690	0.825	0.584
14. The child's pronunciation, articulation and intonation make the child's speech intelligible, despite a possible accent	0.040	0.197	0.764	0.862	0.589
15. The child conjugates verbs correctly	0.013	0.186	0.800	0.894	0.534
16. The child uses (combinations with) nouns correctly	0.006	0.157	0.837	0.916	0.532
17. The child generally constructs correct simple, complex and compound sentences	0.016	0.348	0.636	0.810	0.617

present or not. The right-hand side of Fig. 19.2 shows the empirical cumulative distributions for the different group members. The cumulative distributions were not exactly the same but in light of the sample sizes and the unsystematic ordering of the distributions there was also no reason to conclude that children were disadvantaged if they were player two, three or four. A multilevel regression analysis with children nested in schools, score as dependent variable and group member number as predictor confirmed this conclusion: (member 2-1) $\beta = -0.757$, $z = -1.260$; (member 3-

1) $\beta = -0.272$, $z = -0.440$; and (member 4-1) $\beta = 0.699$, $z = -1.040$. Also, the four analyses at item level showed similar results for the four group members. For example, the p-values differed only 0.052 points on average and the r_{it}-values maximally differed 0.160. A three-level regression analysis without predictors further showed that children within groups were more similar than children across groups. The proportion of explained variance at group level was 0.127. A group effect was thus present in the data, and therefore, there may be occasion to account for group in some analyses. That was not well possible in the present study, however, due to the very small number of groups per school.

Dimensionality was investigated next by presenting the polychoric inter-item correlations in a correlogram. In Fig. 19.3, all correlations are represented by means of a color: darker blue means a higher positive correlation and darker red means a larger negative correlation. As can be seen, all items were positively correlated to each other. The theoretical dimensions of spoken language, however, cannot easily be found. A parallel analysis was therefore conducted in order to determine the number of factors to retain from factor analysis. The results showed that three factors should be retained. This suggestion was adopted. The rotated (pattern) matrix with loadings below 0.300 suppressed is reported in Table 19.5. H2 and U2 represent the communality and specific variance, respectively, of the standardized loadings obtained from the correlation matrix. The communalities were at or above 0.400, except for one just below that value, indicating shared variance with other items. The primary factor loadings were generally above 0.600 and the gap between primary factor loadings and each of the cross-loadings was almost always at least 0.200. Almost no cross-loading was above 0.300, moreover, further indicating that the structure with three underlying factors has a satisfactory fit. Together, the three factors explained 68% of the variance in the items, with factors 1–3 contributing 27, 21 and 20%, respectively. The three factors are in keeping with the three theoretical dimensions of spoken interaction, namely Language form, Language usage and Language content. Some complex cross-factor loadings were nevertheless also present. Especially items 4 (The child integrates contributions from the group into his own contribution when relevant) and 9 (The child shows adequate active listening behavior) did not contribute to one specific factor. Whereas these items theoretically most likely appeal to the social component of conversations, the factor analysis clearly suggested that these items also appeal to substantive quality.

Finally, the reliability and quality of the rating scores was examined. The Greatest Lower Bound was equal to 0.952 and Guttman's Lambda2 was equal to 0.899. These values indicate a very high reliability, but with these values it is not guaranteed that the assessments were also adequate. Therefore, the quality of the rating scores was examined next by comparing the test leader rating scores to an expert rating. Figure 19.4 shows the Mean Absolute Error (MAE) for each test leader. The bottom grey dotted line is the median of the MAE's for the test leaders.The top grey dotted line is the median of the MAE's for an infinitely large number of random assessments. As can be seen, the rating scores for test leaders 1, 4 and 9 were very similar to the ratings scores of the subject-area experts. The rating scores of test leaders 2, 6 and 12, on the other hand, were quite different from the expert rating scores. The MAE

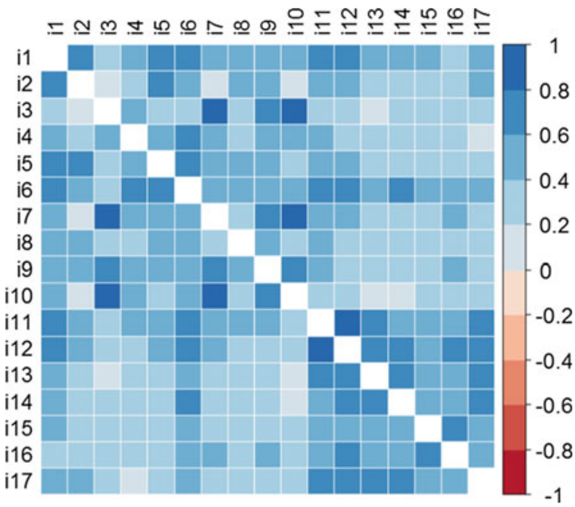


Fig. 19.3 Correlogram of the matrix with polychoric inter-item correlations

Table 19.5 Results three factor principal axis factor analysis

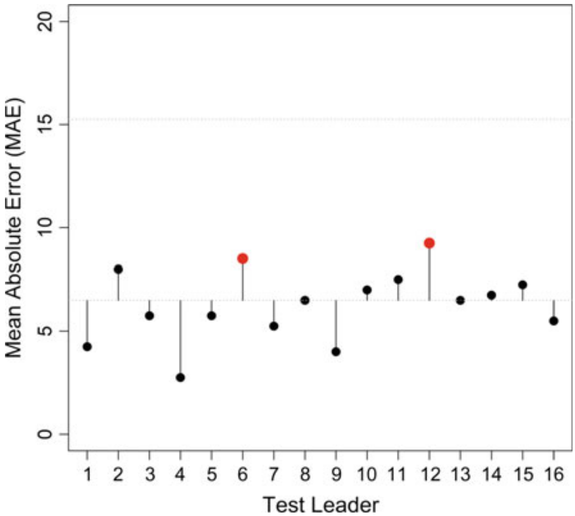
Item	Factor 1: Language form	Factor 2: Language usage	Factor 3: Language content	H2	U2
3. The child usually takes the floor in an appropriate way		0.910		0.870	0.130
4. The child integrates contributions from the group into his own contribution when relevant		0.470	0.450	0.450	0.550
7. The child consistently uses language that fits the situation		0.840		0.810	0.190
9. The child shows adequate active listening behavior		0.610	0.530	0.700	0.300
10. The child consistently gives appropriate verbal and nonverbal responses		0.900		0.860	0.140
1. The child's conversations with the group are meaningful and relevant.	0.440		0.660	0.700	0.300
2. The child regularly takes the initiative to start, continue or stop a conversation			0.830	0.760	0.240
5. The child takes the initiative to achieve a joint communication goal by involving the group in the conversation			0.840	0.790	0.210
6. The child makes his way of thinking understandable.	0.490		0.620	0.710	0.290
8. The non-verbal behavior of the child strengthens his verbal message			0.550	0.390	0.610

(continued)

Table 19.5 (continued)

Item	Factor 1: Language form	Factor 2: Language usage	Factor 3: Language content	H2	U2
11. The child's contribution shows sufficient variation in word use	0.660		0.480	0.720	0.280
12. The child's vocabulary is sufficient to hold a conversation	0.810		0.340	0.800	0.200
13. The child speaks fairly fluently with only occasional hitch, false starts or reformulation	0.800			0.680	0.320
14. The student's pronunciation, articulation and intonation make the student's speech intelligible, despite a possible accent	0.770			0.660	0.340
15. The child conjugates verbs correctly	0.700			0.540	0.460
16. The child uses (combinations with) nouns correctly	0.740			0.630	0.370
17. The child generally constructs correct simple, complex and compound sentences	0.720			0.620	0.380

Fig. 19.4 Mean absolute error per test leader



for test leaders 6 and 12 was even so large that their rating behavior can be considered aberrant in comparison with the other 12 test leaders; the MAD for these test leaders was larger than 2.5. Cohen's weighted kappa coefficient (κ) and Gower's similarity coefficient (G_{xy}) together indicated a fair overall rating quality: $\kappa = 0.307$ and $G_{xy} = 0.815$, where absolute agreement is remarkably higher than relative agreement. On average it did not statistically matter whether the assessment was done by a test leader ($M = 27.078$, $SD = 4.487$) or a subject-area expert ($M = 26.328$, $SD = 5.252$); $t(126) = -0.869$, $p = 0.387$.

19.4 Conclusions and Discussion

In the present study, the *Fischerspiel* board game was used as an entertaining, non-threatening means to evoke conversations between children in special elementary education. Spoken interaction was observed and rated using a newly developed observation form with seventeen performance aspects. It was first examined whether the conversations during the game were sufficiently varied to assess the children's spoken interaction skills. In addition, it was examined whether particular characteristics of the board game were a source of invalidity. When the board game would have elicited very limited or only highly similar conversations, irrespective of the children's skills level, the assessment would, for instance, fail to reveal the differences in skill between children. Sufficient variation was present, however, and the different performance indicators also turned out to function well. The *p*-values were in an acceptable range and all performance indicators had a good discrimination. The performance indicators thus discerned well between children with poor spoken interaction skills and children with good spoken interaction skills. It can therefore be concluded that the board game elicited varied conversations between children and that all aspects of basic spoken interaction proficiency (1F) were observable and assessable. Thus, we can conclude that the assessment did not suffer from underrepresentation of the target skill.

Whether the board game imposed limits that cause (skill) irrelevant variance was evaluated next. Turn taking is one potential source of irrelevant variance, as it might cause differences between children even if their true performance level is equal, but in this study, the order in which children took turns (i.e., first, second, third or fourth in the row) did not significantly affect performance. However, a group effect was found. Analyses showed that the children's performance within groups was more similar than the children's performance across groups. This finding is consistent with several studies on paired test settings where a so-called interlocutor effect was evidenced quite often. Many studies, that is, reported that low skilled children performed better when paired with high skilled children (see, for example, IlSun 2017). More research is needed to study whether the ability of the group members indeed causes differences. At the same time, one should be cautious in using a single paired or group setting in (high-stake) assessments for individual decisions. In high-stakes assessment each child should preferably play with different groups during the observation. For a survey, a group effect might be less problematic as findings are aggregated across groups.

After studying the characteristics of the game the quality of the observation form and the influence of the test leader was considered. The performance aspects provided reliable scores, but the high reliability might in part be caused by the aforementioned group effect. It might be that the test leaders were not able to observe differences between the children within one group really well. A high similarity of the evaluations within groups automatically yields a higher reliability coefficient. As could be expected from theory (see Bloom and Lahey 1978; Lahey 1988), the performance aspects did appeal to three different dimensions: Language form, Language usage and Language content. The first dimension contained all performance aspects that related to grammar, intelligibility and vocabulary. The second dimension contained the performance aspects that related to interaction and alignment on conversation partners. The third dimension contained the performance aspects that related to the quality of the conversation. Finally, the agreement between the expert rating scores and the test leader rating scores turned out to be reasonable. However, two out of sixteen test leaders displayed evaluations that were quite different from the experts' evaluations. The used methodology allows for an early detection of aberrant behavior and with such a methodology a timely intervention is also possible. The test leader could receive extra training, for instance, or some evaluations could be conducted again. This, however was not feasible in the present study, as the experts evaluated the children's spoken interaction skills afterwards from videos. A computer-based assessment in which the test leader and expert can do the evaluation at the same time would speed up the process and potentially prevent test leader effects.

To conclude, the *Fischerspiel* board game proved to be a promising entertaining and non-threatening way of assessing children's spoken interaction skills. Play is important for learning (Mellou 1994), play can be used for learning (so-called serious games, Abt 1970), and play is informative about learning (Otsuka and Jay 2017). Special need children were the target group in the present study, and given the learning obstacles these children encounter, it was crucial to develop an assessment that was practical, short and varied, and would give a feeling of success. The *Fischerspiel* met these criteria, but clearly, also children in regular education could use an assessment with such characteristics. The application of the game as an assessment instrument should therefore also be studied in regular education. Problems associated with observation should then receive particular attention. A computer or online version of the *Fischerspiel* board game might help to overcome some problems. It is then easier to have children playing the game in different groups, the observation can be conducted more unobtrusively and aberrant rating behavior can be detected much faster. Another potential advantage is that in computer games automatic speech recognition technology (ASR) might be used to aid the evaluation. For instance, Ganzeboom et al. (2016) recently developed a serious ASR-based game for speech quality. Such developments are very promising and should therefore certainly be considered when further developing (observer-bias free) assessments for spoken interaction skills. Until then, games like the *Fischerspiel* are a nice alternative.

References

Abt, C. (1970). *Serious games*. New York: Viking Press.

Bloom, L., & Lahey, M. (1978). *Language development and language disorders*. New York: Wiley.

Chapelle, C. A. (1998). Multimedia CALL: Lessons to be learned from research on instructed SLA. *Language Learning & Technology, 2,* 22–34.

Chow, J. C., & Jacobs, M. (2016). The role of language in fraction performance: A synthesis of literature. *Learning and Individual Differences, 47,* 252–257.

Chow, J. C., & Wehby, J. H. (2018). Associations between language and problem behavior: A systematic review and correlational meta-analysis. *Educational Psychology Review, 30,* 61–82.

Cito. (2010). *Checklist toetsconstructie Speciaal (basis)onderwijs*. Arnhem: Cito.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Dickinson, D., Golinkoff, R. M., & Hirsh-Pasek, K. (2010). Speaking out for language: Why language is central to reading development. *Educational Researcher, 4,* 305–310.

Dougherty, C. (2014). Starting off strong: The importance of early learning. *American Educator, 38,* 14–18.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NY: Prentice Hall.

Feldt, L. S. (1993). The relationship between the distribution of item difficulties and test reliabilility. *Applied Psychological Measurement, 6,* 37–49.

Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology, 97,* 493–513.

Ganzeboom, M., Yılmaz, E., Cucchiarini, C. & Strik, H. (2016). An ASR-based interactive game for speech therapy. In *7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, San Francisco, CA, USA, Sept 2016.

Hart, B., & Risley, T. R. (2003). The early catastrophe: The 30 million word gap. *American Educator, 27,* 4–9.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179–185.

IlSun, H. (2017). The effects of paired partners' proficiency levels on test-takers' speaking test performance. *Journal of Research in Curriculum & Instruction, 21*(2), 156–169.

Kent, S., Wanzek, J., Petscher, Y., Al Otaiba, S., & Kim, Y. (2014). Writing fluency and quality in kindergarten and first grade: The role of attention, reading, transcription, and oral language. *Reading and Writing: An Interdisciplinary Journal, 27,* 1163–1188.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal, 70,* 366–372.

Lahey, M. (1988). *Language disorders and language development*. New York: MacMillan.

Landers, R. N. (2014). Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation & Gaming, 45,* 752–768.

Ledoux, G., Roeleveld, J., Langen, A., & van Smeets, E. (2012). *Cool Speciaal. Inhoudelijk rapport* (Rapport 884, projectnummer 20476). Amsterdam: Kohnstamm Instituut.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49,* 764–766.

McLaughlin, S. (1998). *Introduction to language development*. Londen: Singular Publishing Group.

Meijerink, (2009). *Referentiekader taal en rekenen*. Enschede: SLO.

Mellou, E. (1994). Play theories: A contemporary review. *Early Child Development and Care, 102*(1), 91–100.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry to score meaning. *American Psychologist, 50,* 741–749.

Michaels, J. (1983). Systematic observation as a measurement strategy. *Sociological Focus, 16*(3), 217–226.

Nation, K., & Snowling, M. J. (2004). Beyond phonological skills: Broader language skills contribute to the development of reading. *Journal of Research in Reading, 27,* 342–356.

Naucler, K., & Magnusson, E. (2002). How do preschool language problems affect language abilities in adolescence? In F. Windsor & M. L. Kelly (Eds.), *Investigations in clinical phonetics and linguistics* (pp. 99–114). Mahwah, NJ: Erlbaum.

O'Malley, J. M., & Pierce, L. V. (1996). *Authentic assessment for English language learners: Practical approaches for teachers*. New York: Addison-Wesley.

Otsuka, K., & Jay, T. (2017). Understanding and supporting block play: Video observation research on preschoolers' block play to identify features associated with the development of abstract thinking. *Early Child Development and Care, 187*(5–6), 990–1003.

Shanahan, T. (2006). Relations among oral language, reading and writing development. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 171–183). New York: The Guilford Press.

Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. Cambridge, MA: The MIT Press.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107–120.

Stieger, S., & Reips, U. D. (2010). What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior, 26,* 1488–1495.

Taylor, L., & Galaczi, E. (2011). Scoring validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking*. Cambridge, UK: Cambridge University Press.

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69,* 613–625.

Van Langen, A., Van Druten-Frietman, L., Wolbers, M., Teunissen, C., Strating, H., Dood, C., et al. (2017). *Peilingsonderzoek Mondelinge Taalvaardigheid in het basisonderwijs*. Nijmegen: KBA.