3. Method

3.1. The base study: BeLesen

This study is one component of an ongoing four year longitudinal study entitled "BeLesen: The Berlin Longitudinal Study of Reading Competence Development among Primary School Children" led by Professor Hans Merkens at the Institute for Educational Science at the Free University of Berlin from 2002-2006. The project was initiated to examine the differential effects of various reading instruction methods on the acquisition of reading and subsequent academic performance of children from Turkish families compared to native German-speaking children. BeLesen encompasses 59 classes in 30 schools located in the lowest SES districts in central Berlin. At the onset of the study, the 1237 children in the sample ranged in age from 4.8 to 9.3 years, (M = 6.9 years). Of the minority language children in the sample, over 90% had attended preschool or kindergarten before beginning the first grade, and can thus be considered participants in the German school system from the onset of their school careers. In other words, the vast majority of children from the minority language households have had their entire schooling experience in Germany, not in their parents' country of origin.

3.2. Participants

3.2.1. Sample description

Children were selected from the existing pool of 1237 first grade children participating in the longitudinal study described above[1]. To set recruitment goals, an assessment of the required statistical power was carried out with the help of Cohen's (1992) table of suggested sample size for .80 power. In this case, power refers to the probability of detecting a true effect or significant difference and is equal to 1 – the probability of a Type II error, or 1 – ß (Welkowitz, Ewen, & Cohen, 2000). According to Cohen, the .80 specification for power (ß = .20) is a convention that has been proposed for general use (see also High, 2000): "A materially smaller value than .80 would incur too great a risk of a Type II error. A materially larger value would result in a demand for N that is likely to exceed the investigator's resources" (p. 158). In effect, Cohen's power

[1] Neither the larger BeLesen study nor the sub-sample used for the present study aim to be representative samples for German primary school children. These are non-representative samples intentionally collected in socio-economically disadvantaged areas in inner-city Berlin with high proportions of immigrant families. In light of the large and growing number of children living in similar urban areas in Germany, it is essential to investigate this population regardless of far-reaching generalizability. Nonetheless, the current sample can be assumed to be generalizable to other Turkish-German inner-city populations.

analysis was used to anticipate the likelihood that the study could achieve a significant effect by finding a sample size that balances the risk of making a Type I or Type II error.

The posited hypotheses outlined in Section 2.2. and reviewed literature led to the expectation that mid-sized effects could be anticipated. Based on the medium-sized effects anticipated for the planned mean difference analyses (*d* = .50 according to Cohen's Table 1 in, p. 157), it was determined that at least 64 participants would be required for each group to test at a .05 level. The expected medium effect sizes for the multiple regression analyses (*d* = .15, according to Cohen's Table 1, p. 157) led to a recruitment goal of at least 91 for five independent variables also based on a .05 significance level.

Fourteen of the 59 classes involved in the larger study were selected on the basis of similar average cognitive abilities test scores across classes, class composition of at least 50% minority language students, and location in school districts of similar socio-economic status (SES). This was done to avoid data collection in classes with significantly greater socio-economic advantage or classes with a disproportionate number of children with unusually high or low cognitive test scores. In order to address the fact that cognitive skills of bilinguals are usually underestimated by standardized intelligence tests (Cline & Frederickson, 1999), scores were transformed into *z*-values for the German and the Turkish children before averaging and comparing class scores. Only classes with similar average class scores were selected to participate. All six schools in which the 14 classes were embedded fell into SES zones with scores of 6 or 7 (1 being a highly advantageous district and 7 being highly disadvantageous). As can be seen in Table 2, the 14 teachers had many years of teaching experience (*M* = 22.92 years), similar class sizes, and the classes had a substantially larger proportion of minority language students than native German speakers (*M* = 67%).

As many children as possible were tested in each class, although not all could be considered in the final analyses (prerequisites for inclusion in the final analyses are described below). Initially, 251 children were tested in the first wave of oral data collection in the middle of second grade (T1). The total number of children present and able to be tested at both central data collection points for this investigation, in the middle (T1) and at the end of second grade (T2), was 224.

Table 2

Descriptive Characteristics of the 14 Participating Classes

Class number	Teacher's years of experience	Class size	Percentage minority language students	Berlin district	School district SES	*N* of participants used in final analysis
1	13	21	81	Neukölln	7	12
2	26	21	79	Neukölln	7	11
3	30	23	65	Neukölln	7	14
4	20	24	63	Neukölln	7	16
5	19	25	60	Neukölln	7	15
6	27	25	64	Kreuzberg	7	9
7	28	25	68	Kreuzberg	7	9
8	19	25	80	Kreuzberg	7	12
9	--	21	76	Kreuzberg	7	13
10	26	13[a]	69	Kreuzberg	7	4
11	26	24	58	Schöneberg	6	12
12	24	23	61	Schöneberg	6	18
13	17	25	76	Kreuzberg	6	17
14	--	21	50	Kreuzberg	6	7

Note. The two teachers who did not provide their years of experience are marked with a dash (--).

[a] Class 10 was particularly small since it was constructed as a "Förderklasse" for providing additional attention to at-risk or minority language children.

Inclusion into one of the two language groups (Turkish-German bilingual, TB; German monolingual, GM) was determined by three factors: child self-report of home language use over two times of measurement, a teacher report of the children's home language at two separate times of measurement (T-1 and T1), and for the Turkish-speaking children, verbal Turkish assessments over two points of measurement. Table 3 shows the number of participants excluded for each criterion. Because it was not possible to test the home language skills of each language spoken by children in the sample and thus impossible to determine their level of bilingualism, participants who reported speaking a language at home other than German or Turkish were removed from the sample, thus excluding the possibility of including a further comparison group. Participants in the native German-speaking group were admitted to the study only if German was reported to be the only language spoken by both parents in the home. These stipulations resulted in the removal of 53 children whose bilingualism or monolingualism could not be ensured due to conflicting reports or exposure to languages other than Turkish or German in the home.

In order to ensure a base level of bilingualism, children in the Turkish bilingual group were required to demonstrate a base level of Turkish over two points of measurement falling no more than one standard deviation (*SD*) below the average of the reportedly Turkish-speaking sample. The operational definition for Turkish/German bilingualism in this study involves average proficiency in verbal Turkish skills for an immigrant population, a minimum of two years of schooling in German, both self and teacher reporting of a home language other than German, and sufficient German skills as demonstrated in school achievement tests. All children in the study had attended German school from the onset of their education and were proficient enough in German to complete all German language tests. There were, however, participants who reported some Turkish spoken in the home yet, scored very low Turkish verbal abilities scores. They were eliminated from the analyses to avoid confounding the bilingual sample with non-bilingual children from Turkish families. This resulted in the removal of two cases from the final analyses, thus bringing the total number of children in the bilingual group to 100.

Table 3

Exclusion Criteria and Sample Sizes

	Exclusion Criteria			
	Initial categorization	Non-Turkish or inconclusive home language reports	Insufficient performance on Turkish language assessment	Total *N*
Bilingual	151	49	2	100
Monolingual	73	4	--	69

Although there is no single definition for when bilingualism is achieved, many studies consider a child to be bilingual if it is raised in a household where a language other than that of the surrounding community is spoken (e.g., Queen, 2001). The operational definition used here could therefore be considered relatively stringent. The strict inclusion criteria were chosen for this study since several of the hypotheses are dependent on the participants' dual language abilities and proficiency in Turkish. It was also considered important to avoid making assumptions about the participants' language abilities based purely on expectations drawn from the parent's cultural heritage.

Table 4 illustrates the self-reported language behaviors of the Turkish bilingual participants. In an interview at the end of second grade, the children were asked to rate how much they spoke Turkish with various members of their families and the amount of Turkish language television they watched (0=*never*, 1=*a little*, 2=*often*, or 3=*almost always*). Almost all children (96%) spoke some amount of Turkish with their mothers or fathers but fewer children reported speaking Turkish with their siblings or friends. Since Turkish was spoken with parents "often" on average (*M* = 1.99 for mothers, *M* = 1.90 for fathers) but less frequently than with other family members, it can be readily assumed that German is a substantial language of communication in the household. Although Turkish television is available in most Turkish homes via satellite, and a majority of the children reported watching some Turkish television, they report doing this very infrequently. Again, it could therefore be assumed that a substantial amount of German language television is being watched in most homes of participants in this study. It is possible that this self-reported survey of language practice may not be reliable among children of such a young age, but since information from additional sources was not available, it was deemed useful to at least consider the information provided by the children, who indicated that a considerable amount of mixed German and Turkish language was used in the home. It is important to keep this in mind for understanding the Turkish bilingual participants' linguistic experiences and environment.

Table 4

Turkish Language use by Participants in the Turkish Bilingual (TB) Group, Including Percentage of the TB Group who Report Speaking Any Amount of Turkish with each Family Member and Mean Reported Amount of use (0=Never or NA, 1= A Little 2 = Often, 3 = Almost Always) with Standard Deviations (SD)

	Percentage	*M*	*SD*
Mother	96	1.99	0.95
Father	96	1.90	0.90
Siblings	70	1.07	0.95
Friends	67	0.75	0.64
Other people living in home[a]	14	0.33	0.89
Television	73	1.03	0.86

[a] Refers to entire sample, not only those who reported additional people in living in the household

After removing participants who did not fit the required language profiles, there were 69 children in the German monolingual group and 100 children in the Turkish-German bilingual group. Participants were between the ages of 7.1 and 9.5 (*M* = 8.0, *SD* = 0.4) at the onset of the verbal data collection at T1 and had just begun 2nd grade. Background data had been collected with the larger sample in which the current sub-sample was embedded since the beginning of first grade. None of the participants for this study had repeated the first, second, or third grade. The sample was composed of 55.1% females distributed evenly across the bilingual and monolingual groups. Analyses of variance and chi-square analyses found no significant differences between the two groups with regard to age, school district SES[2], sibling ages, sex, or country of birth (see Table 5). Table 5 shows the results of an ANOVA indicating no group differences in cognitive abilities in the first grade (T-1).

Table 5

Means and Standard Deviations for Demographic Variables and Comparisons (ANOVA and Chi-square) of Both Groups at T1

		GM	TB	F	*p*
Age at T1		7.92	7.92	0.00	.99
	SD	*0.44*	*0.38*		
School district SES		6.62	6.72	1.76	.19
	SD	*0.49*	*0.45*		
Siblings					
Total number of siblings		1.45	1.65	0.94	.33
	SD	*1.23*	*1.34*		
Number of older siblings		1.00	1.04	0.06	.81
	SD	*1.07*	*1.07*		
				Pearson's χ^2	*p*
Sex:					
Percentage female		55.1%	49.0%	0.60	.53
Country of Birth:					
Percentage outside of Germany[a]		4.3%[b]	10.0%	1.84	.24

Note. GM = German Monolingual, TB = Turkish Bilingual. Socio-economic status (SES) of school districts is based on a composite score from the local Berlin municipality (1 = Highly advantageous, 7 = Highly disadvantageous).
[a] Child self-report. [b] One monolingual German-speaking girl reported being born outside of the EU, one monolingual German-speaking boy reported being born in an EU country outside of Germany.

Table 6

ANOVA with Means and Standard Deviations for Both Groups on Cognitive Abilities at T-1

		GM	TB	F	*p*
Base cognitive skills at T-1 (raw score max. 36 pts.)		25.39	25.07	0.13	.79
	SD	*5.53*	*5.02*		

Note. GM = German Monolingual, TB = Turkish Bilingual.

Although no group differences were found with regard to demographic characteristics or cognitive abilities at the onset of the study, the groups did differ on the reported amount of parental literacy behaviors in the home (see Table 7). As described in Section 1.2.6., it is

[2] Due to the structural framework of the BeLesen study, it was not possible to collect SES-related information from participants' parents. To compensate for the lacking parental data, an instrument measuring cultural capital was administered to the participants as an indicator of SES. This instrument, Books in Home, will be described in further detail in the Measures section of this chapter.

internationally well documented that parent reading behaviors and pre-primary education play an important role in the development of language skills, particularly for minority language and lower income children (e.g., Grimley & Bennet, 2000; Golova et al., 1999; High et al., 2000; Spiess et al., 2003). It was therefore considered important to take these "non-academic" influences on language and literacy into consideration. Parental reading aloud practices were investigated with the Home Language Interview at T1. Children were asked, "How often do your parents read to you?" and their responses were coded as daily, sometimes, or rarely. It was found that, although a similar proportion of monolingual and bilingual children reported being read to very infrequently (around 50% of each group), the small number of children who reported being read to daily was greater among the German monolingual children (TB = 3.1% compared to GM =14.5%) were read allowed to on an every day basis, χ^2 (2, N=166) = 7.26, p = .03. Since the singing of children's songs and reciting of rhymes are thought to be strongly related to phonological awareness, participants were also asked to rate how often their parents engaged in these activities with them. No differences were found between the two groups on the frequency of such singing and rhyming behaviors. In fact, very few families in either group engaged in such activities.

Table 7

Chi-square Analysis of Home Literacy-Related Experiences and Childcare

			GM	TB	Pearson's χ^2	p
Literacy behaviors of parents						
	Reading aloud					
		Daily	14.5%	3.1%	7.26	.03
		Sometimes	38.6%	44.3%		
		Rarely	46.6%	52.6%		
	Singing or reciting rhymes and poems[a]		30.4%	25.0%	0.61	.48
Childcare experiences						
	Preschool attendance[b]		98.3%	97.8%	1.05	.59
	After-school daycare attendance[c]		66.7%	52.0%	3.60	.04

Note. GM = German Monolingual, TB = Turkish Bilingual.

[a] Yes or no response format. [b] Reported by the classroom teacher. [c] Reported by the child.

Table 7 also shows a comparison of the two groups on the participants' childcare experiences. According to the teachers' reports, around 98% of children in both groups attended preschool or a daycare program. However, significantly fewer children in the bilingual group reported attending after-school childcare (52%) than their monolingual counterparts (67%) at the time of the individual verbal testing, χ^2 (1, N=169) = 3.60, p = .04.

3.2.2. Test administration

Schools in the larger BeLesen study were recruited through cooperation with the Berlin Ministry of Education and encouraged by the Ministry to participate. For the sub-sample used in this study, teachers in districts with large minority populations and low SES were contacted and asked for permission to take small groups children from class for 45 minutes during the school day as a part of the BeLesen investigation. Since the teachers had already been involved in project testing for a year, all readily agreed to allow their students to participate in the individualized verbal testing for this study.

The written measures of reading and cognition were administered by trained masters-level university students in group tests as a part of the BeLesen study. These tests were administered in 6 month intervals over a period of two days (2 hours per day). As is typically recommended for children of this age group, testing was conducted in the mornings after the first hour of school with the aim of accessing a higher energy phase of the day for the children. Phonological awareness, vocabulary, and memory were assessed by the primary investigator and a team of trained Turkish-German bilingual graduate students. The verbal assessment battery that was administered at half-year intervals in the second grade lasted around 45 minutes. Children were rewarded with stickers for the individual testing and seemed to find the oral tasks innately enjoyable and rewarding. Neither participant motivation nor teacher cooperation was problematic.

3.3. Materials

Three types of measures were included in the present investigation. *Baseline and background* measures include a teacher assessment (including evaluations of each individual participant's language skills, perceived home language practices, and school behavior), a test of cognitive abilities, a children's self-report measure of the number of books in the home, and a home language practices interview. The *literacy performance measures*, administered in classrooms longitudinally as part of the larger BeLesen study, include measures of word decoding and reading comprehension. All BeLesen literacy and cognitive skills instruments required written responses only and were administered in group form. The third category of instruments encompasses the individually administered *verbal measures* for testing phonological awareness, verbal abilities in Turkish and German, and verbal short-term memory. A detailed timetable (Table 10) at the end of this section provides an overview of which instruments were administered at each point of measurement. For further reference, unpublished and modified measures that would otherwise be unavailable have been included in Appendix A.

3.3.1. Baseline and background measures

Teacher Assessments of Language and Learning (T-1, T1)

This assessment scale is a subjective measure provided by the classroom teacher regarding the language and learning abilities of individual children at the beginning of the first grade and in the middle of second grade. The *Teacher Assessments of Language and Learning* measured three individual traits: *German Language Abilities*, *Readiness to Learn*, and *Concentration.* All items in the teacher assessment scales were administered with four level response scales (1= *low* to 4 = *high*). The psychometric characteristics provided below are based on the entire BeLesen sample of approximately 1200 children[3].

The six item *German Language Abilities* scale first administered in November of 2002 asked teachers retrospective questions regarding the children's abilities at the beginning of the school year (August 2002). Items included questions such as, "Could the child express himself appropriately in German? " and utilized four-point response scales (1 = fully insufficient to 4 = fully sufficient). The scale demonstrated overall high reliability (alpha =.97) at T-1 and satisfactory reliability at T1 (alpha = .87) with the full BeLesen sample of 1216 participants. External validity is indicated by moderately high correlations (*r* = .45 - .57) with the German verbal abilities measures taken at T1 and T2 (see Table 22 in the Results section).

The four-item *Readiness to Learn* scale aimed to assess classroom behaviors that indicate motivation. It included items such as "participates in lessons" (four point response scale: 1=*never* to 4 = *always*) and demonstrated similarly sufficient reliability at both T-1 and T1 (T-1 alpha = .87, T1 alpha = .86). Teachers were also provided with the opportunity to rate their students on their ability to concentrate by means of a three item scale with the same response options as the *Readiness to Learn* scale (e.g., "is easily distracted"). Comprised of three items, the *Concentration* scale's reliability was also satisfactory with an alpha of .83 at T-1 and .84 at T1. A copy of the instrument administered at T-1 is included in Appendix A.

Separate calculations of reliability for children in the Turkish-German group and the German group showed no differences in alpha coefficients (see Table 8). At T-1, there were no differences in reliability between the two groups on the German Language scale or on the

[3] Since the Turkish heritage children in the BeLesen sample were not tested for language abilities, they are not referred to here as "bilingual" as is the case in the majority of the following analyses conducted with the language-tested sub-sample of bilingual and monolingual children. Instead, the minority language children from Turkish backgrounds are referred to as "Turkish-German", and the children from assumed German only families as "German".

Readiness to Learn scale, and a negligible difference between the groups on the Concentration scale (German α= .85; Turkish-German, α = .83). Finally, teachers were asked to provide information regarding the student's ethnic backgrounds and home languages. This information was compared with the children's reports of home language use for the purposes of determining group inclusion or exclusion as described in section 3.2.1..

Table 8

Differentiated Reliability Coefficients for the German and Turkish Groups on the Teacher Assessment Scales at the Beginning of 1st Grade (T-1) and in the middle of 2nd Grade (T1)

	T-1		T1	
	German	Turkish-German	German	Turkish-German
German Language	.83	.83	.83	.82
Readiness to Learn	.87	.87	.83	.87
Concentration	.85	.83	.85	.83

To gain further insight into the psychometric constructs of the teacher assessment instruments, the intercorrelations of the teacher assessment instruments scales were also explored (Table 9). It appeared that the scales were relatively stable over time from early Grade 1 to mid Grade 2 (r = .50 - .66). Table 9 also shows that while the Readiness to Learn scale correlated strongly with the Concentration scale (r = .37 - .62), the German language scale was essentially unrelated to concentration abilities (r = .12 - .24).

Table 9

Scale Intercorrelations for the Teacher Assessment Measures at T-1 and T1

	1	2	3	4	5	6
1 German language abilities beginning of Grade 1	--					
2 German language abilities middle of Grade 2	.63**	--				
3 Readiness to learn beginning of Grade 1	.34**	.20*	--			
4 Readiness to learn middle of Grade 2	.23*	.34**	.50**	--		
5 Concentration abilities beginning of Grade 1	.12	.14	.61**	.37**	--	
6 Concentration abilities middle of Grade 2	.13	.24**	.49**	.62**	.66**	--

Note. Using a Bonferroni approach to control for Type 1 errors across the 6 correlations results in a *p*-value cutoff of .008 to indicate a significant correlation.
**p* < .05. ***p* < .008 (Bonferroni cutoff).

Cognitive abilities: Culture Fair Intelligence Test 1 (T-1)

Three subtests of the Culture Fair Intelligence Test (CFT1; Cattell, Weiß, & Osterland, 1997) were administered to measure fundamental non-verbal cognitive skills and general fluid abilities in the first half of the first grade. Subtests 2-4 (*Visual Processing*, C*lassification Skills*, and *Detail Recognition*) with 36 possible points were administered at T-1 in November 2002[4]. Test duration was approximately 10-15 minutes. The internal validity of the CFT1 is considered to be very good, with reasonable item-total correlations and sound factor analytical composition.

For the Berlin sample of inner-city children (N = 1234), reliability was good (α = .85) for the three subtests together and had a similar reliability coefficient to the national standardization sample (N = 1710) of first graders (α = .90) for the entire measure. However, presumably, since Subtests 1 and 2 are timed measures, the authors of the CFT did not calculate Cronbach's alpha to determine the internal consistency. Therefore, no separate reliability information was available for the national standardization sample for the first two subtests.

Although the CFT is designed to be a culture free test of cognitive abilities independent of verbal abilities in the test language, it has often been reported that L2 language learners are underestimated in intelligence tests administered in their L2 (for an overview see Cline, 2000). However, data from both the CFT standardization sample and from the larger BeLesen study indicate that the CFT is an acceptable measure of cognitive abilities for the minority language sample at hand. As part of the national standardization procedure for the German CFT, 75 minority language children were compared with 75 native German-speaking children. No mean differences in scores were found between the two groups. In order to examine possible differences in the psychometric properties of the CFT for German and Turkish-German participants, separate reliability analyses for the two groups were conducted with participants from the BeLesen sample. Test reliability for the Turkish-German children in the sample was very similar to that of the German monolingual children (German: α = .78; Turkish-German: α = .84). In spite of these positive indicators, as with all cognitive measures administered to non-native speakers, the CFT should be viewed with a certain amount of caution for describing the present Turkish-German population.

Measure of Home Literacy (T1, T2)

This single-item pictorial self-report measure for children of *Books in Home* was taken from the ELEMENT study conducted by Lehmann & Nikolova (2005). This instrument is an

[4] The CFT offers the user three tables of standardized scores, one for the sum of subtests 1-2, one for the sum of subtests 3-5, and one total summed score for all five subtests. Since the BeLesen investigation utilized only scales 2-4, it is not possible to fully compare the scores of this sample to the national standardization sample.

indicator of familial background and educational resources. Participants are asked to select from a list of five options illustrated with pictures of bookshelves to designate how many books their family has: "none or very few (0-10)" to "enough to fill three or more bookshelves (over 200)".

The indicator of 'books in the home' has proven to be one of the best predictors of academic performance in both German and international investigations (e.g., Lehmann & Nikolova, 2005; Verhoeven & Aarts, 1998). Although not a standardized instrument, the scale was developed on the basis of research findings in this area. It is typically used with children in the fourth grade or beyond. It is therefore not surprising that this self-report measure was not found to be particularly reliable with the second grade children in the current sample. The responses at T1 correlated with responses at T2 with a correlation coefficient of only .49 for the Turkish bilingual children and with $r = .35$ for the German monolingual children. Even though the means remained the same for both groups over T1 and T2, and the groups differed significantly from one another at both times of measurement (the German monolingual group reported having significantly more books at home than the Turkish bilingual group), the low test-retest reliability indicates an inadequate reliability[5]. The measure was therefore unfortunately not utilizable for the following analyses.

Home Language Use Interview (T2)

This is an instrument of home language practices developed particularly for this investigation. The eight-item measure aimed to reveal the languages used in the home and the extent to which the child speaks them. Although the interview was scripted in German, test administrators had the option of asking questions in Turkish if they felt that additional clarification in Turkish would be helpful or put the child more at ease. Turkish was then used in fewer than half of the interviews with Turkish-German bilingual participants.

Participants were asked by the interviewer if a language other than German is spoken in the home and if so, which one. The participants were then asked to report how often (0 = almost never, to 3 = almost always) they speak that language with their mother, father, siblings, additional residents in the home, and how often they watch television programs in that language. This not only allowed for a differentiated picture of what languages are spoken by whom in the home, but also provided a 18 point scale indicating how much the minority language is spoken and heard by the child. A copy of this eight item short interview is found in Appendix A.

[5] In other areas of this study, similar test-retest correlation coefficients will be considered satisfactory. However, since this scale is meant to examine a fact related to SES and not an academic ability, a much higher reliability is required. In a similar vein, the educational level of a parent would also not be expected to show variability over a six-month period.

3.3.2. Literacy performance measures

Word decoding: Würzburger Leise Leseprobe - Würzburg Silent Reading Test, (T0, T1, T2, T3)

The Würzburg Silent Reading Test is an economical speed-based reading test designed to identify word level reading processing speed and word decoding in German (WLLP; Küspert & Schneider, 1998, 2001). This instrument serves as the primary measure of word decoding for this investigation. The WLLP is a paper and pencil test administered in group form with a time limit of five minutes. Participants are presented with series of written words. Each word is followed by four pictures, from which the participant is instructed to select the one that best represents the written word. Participants are instructed to find and mark the appropriate pictures for as many words as possible before the five-minute time limit is reached. A total score is calculated based on the number of pictures correctly identified within the limited time frame. Because this measure is expandable in length, it is well suited for measuring growth in longitudinal studies. At T0 and T1, 80 items were presented, 120 items were presented at T2, and 140 items at T3. Two parallel forms were administered in which the items were the same, albeit in a different order.

The reliability of the WLLP was, in part, determined with the standardization sample by correlating performance on the two parallel forms for children who were given both test versions (Küspert & Schneider, 2001). In the first grade, performance on the two parallel test forms correlated with $r = .87$, with $r = .92$ in the second grade, and with $r = .93$ in the third grade. Test-retest reliability within a 14-week time period was .75 for the first grade, .81 for the second grade, and .88 for the third grade. The WLLP can therefore be considered a reliable instrument for use in the early primary school years. The standardization sample also demonstrated high correlations with other standardized reading measures and teacher evaluations. It is important to mention, however, that the standardization sample excluded children who were second-language learners of German and included no children from Berlin schools. Furthermore, since the BeLesen sample of inner-city children did not administer parallel test forms to the same children, no reliability coefficients are available for this sample. The only available reliability information for the WLLP is therefore based on populations relatively different than the current sample.

A study by Kornmann, Geider, Jester-Zürker, and La Selva (2003), looked more closely at third through eighth grade children from multilingual households who were enrolled in special education schools. They found that home language played a significant role in the comprehension of the WLLP item vocabulary, but did not significantly affect decoding itself. In other words, although non-native speakers had lower levels of comprehension for the items presented in the WLLP, their decoding scores were similar to the those of the native German speakers.

Overall, the test characteristics provided in the WLLP manual should be interpreted with caution when the test is applied to a multilingual sample of children in the traditional Berlin public schools. There is too little information available indicating its validity or reliability in a sample such as that at hand. External validity of the WLLP for this particular sample will be briefly discussed in the Results section below.

Reading Comprehension: Ein Leseverständnis Test für Elementarschüler (ELFE)-- (T2, T3)

The ELFE reading comprehension measure (Lehnart & Schneider, 2005) administered as a group test at T2 and T3 serves as the primary dependent variable for this study. The ELFE was designed to be used with children in the first through sixth grades. For this investigation, only one of the four subscales, *Text Comprehension*, was used. This subtest was developed to test a child's ability to find information in a text, to infer meaning beyond written sentences, and to draw conclusions about that text. A series of short texts (2-3 sentences) are provided in a test booklet, each followed by one or several questions regarding the content of the text. In total, there are 20 questions and thus a maximum of 20 possible points. The administration of this measure lasted under ten minutes including instruction.

The ELFE was still in the process of national standardization at the time of this investgation, but a limited number of provisionary test characteristics are already available from a preliminary standardization sample of $N = 100$ (grades 1-4). Internal consistency values (Cronbach's alpha) for second grade children in the standardization sample was $\alpha = .94$ compared to $\alpha = .84$ for the approximately 1200 second graders (T2) in BeLesen's inner-city sample. For the third grade, the standardization sample had an internal consistency of $\alpha = .96$ compared to $\alpha = .87$ in BeLesen's Berlin sample. Within the BeLesen sample, the ELFE demonstrated slightly higher reliability for the German group (T2, $\alpha = .89$; T3, $\alpha = .90$) compared to the Turkish-German group in the BeLesen sample (T2, $\alpha = .74$; T3, $\alpha = .82$); however, the differences were not significant.

From the preliminary standardization data, a few measures of external validity are available, although not reported separately for minority and majority language students. As a test of external validity, the ELFE Text Comprehension subtest was found to be substantially correlated both with scores from the WLLP ($r = .58$) and teacher assessments of reading ability ($r = .65$ in second grade). These substantial correlations with a different type of reading test can be interpreted as an indication of external validity of the ELFE Text Comprehension measures.

3.3.3. Verbal measures

Measurement considerations for the collection of verbal data

Although a number of investigations in English speaking countries have looked at phonological awareness in kindergarten and preschool aged children as a predictor of later reading skills, this study begins testing for phonological awareness in second grade. A primary reason for initiating phonological testing in the second grade for this study was that, in contrast to English speaking children, research has shown that German-speaking children do not demonstrate measurable levels of phonological awareness before the end of first grade (Landerl et al., 1992). Nonetheless, many studies examining phonological awareness (particularly in conjunction with literacy and reading comprehension) have used samples starting at second grade and above (e.g., McBride-Chang, 1995).

Furthermore, many studies examining the predictive powers of early base reading skills in preschool and kindergarten children have collected information on the child's knowledge of letter and ability to recognize alphabetic characters. Although letter knowledge measured in young children often carries a significant amount of weight in predicting later reading skills (e.g., Muter & Diethelm, 2001), due to anticipated ceiling effects, it was not considered a suitable skill area to measure in the first and second grades and was thus omitted from the current study.

A further deliberation regarding measurement in this study was a general lack of unity in instruments used in studies of phonological awareness. All too rarely are the stimuli used in one study repeated in subsequent investigations. This is particularly the case in research conducted with non-English speaking populations. Not surprisingly, there are no internationally standardized tests that function in a range of languages. This also plays a substantial role in the wide range of phonological awareness tests. A test derived from non-words could be one solution to this dilemma, but the frequency of different consonant and vowel sounds is divergent from language to language. According to McBride-Chang (1995), there are several essential components typically common to instruments measuring phonological awareness in the literature. In order to promote consistency in phonological awareness research, McBride-Chang recommends that the following four aspects of phonological awareness measures are considered:

- **Perception:** Participants listen to one or more orally presented words or non-sense words and are asked to repeat the words to ensure that the stimulus was correctly perceived.
- **Memory:** Participants hold the speech segment in memory long enough for an operation to be performed.
- **Operation:** Participants manipulate sounds (e.g., identifying single phonemes or choosing a word from a list that "does not belong").
- **Oral Communication:** Participants articulate responses orally.

To address these issues of commonality across studies, the measures selected and adapted for the present investigation were chosen and modified based on these four essential components of a phonological awareness instrument. The McBride-Chang concept of phonological awareness is particularly useful because it is relatively simple to operationalize and it is easily applied in designing empirical investigations.

The three levels of phonological awareness (syllable, phoneme, and onset/rime) proposed by Liberman et al. (1974) and Goswami and Bryant (1990) and described in detail in Chapter 1 (p.18), were also taken into account in the selection of suitable measures for this investigation. Since the children in the current sample possessed highly developed verbal systems and were already beginning to learn alphabetic principals, the syllable-level of phonological awareness was deemed too simplistic. Since, it is easily measured in both school-aged children and adults, the second level, involving phonemes was considered more suitable for measuring the phonological awareness of early elementary school children. While the onset/rime level of phonological awareness is a useful concept, there was not sufficient empirical evidence to warrant its use in this study. For reasons of age appropriateness and a clear research base, only instruments measuring phonological abilities on the level of phonemes were selected for use in this investigation.

Several further recommendations were taken into consideration in the selection of measures. Phonological awareness scales were based primarily on pseudoword items whenever possible for several reasons. First, non-word items are thought to encourage purer phonological processing and discourage the use of orthographic strategies (Stuart, 1990). Second, pseudowords were considered particularly advantageous stimuli for studies in multilingual contexts such as this, because they could be tailored for language neutrality. The non-word stimuli were adapted to be pseudowords in both Turkish as well as German, in order to minimize advantage or disadvantage for either group. Finally, the high correlation between phonological awareness in the two languages of bilingual children in the existing literature (e.g., Durgunoğlu et al., 1993; Verhoeven, 1994) was high enough to suggest that phonological awareness was more of a general language-unspecific cognitive ability than a language-dependent skill that differed for both languages.

Finally, a range of difficulties in accessing phonological awareness is used to measure a wide breadth of skills in the participants. As recommended by McBride-Chang (1995), a variety of consonant types are presented (e.g., simpler frictives such as /ff/ or /ss/ are mixed with stop consonants such as /g/ and /k/). However, because consonant clusters within a phoneme (e.g. /tr/ or /pl/) are highly unusual in Turkish (Öney & Durgunoğlu, 1997; see 1.1.2. for a detailed discussion), they were eliminated whenever possible in the phonological awareness measures.

There is less detailed literature on the measurement of verbal memory in combination with reading studies. Because Näslund and Schneider (1991) defined verbal memory as a separate component of the early reading process whereas others defined it as a component of phonological awareness (e.g., Wagner et al., 1993), it is considered both independently and as an aspect of phonological awareness. Therefore, similar guiding principles were applied in the selection of an appropriate verbal memory instrument as for the phonological awareness measures. The instrument was required to be based on pseudowords with sounds that have similar frequency in both Turkish and German and have a wide range of difficulty.

Listening comprehension is often cited as an important component of reading (e.g., Marx, 1998, Proctor et al., 2005) and frequently understood as one component of the verbal abilities skill set. Measures of listening comprehension have therefore been included in this study to diversify the measurement of verbal abilities. Although it is not explicitly named in the Näslund and Schneider (1991) model description as a central aspect of verbal abilities, listening comprehension is explored here as a potential enrichment of the verbal abilities component of the model. To explore its relationship to vocabulary knowledge and reading comprehension, a measure of listening comprehension was included at T2.

Finally, a measure of vocabulary for both Turkish and German was administered to the bilingual children, whereas only German vocabulary tests were administered to German monolinguals. At the time of the study onset, no German standardized instruments existed with which both languages could be measured. For that reason, this investigation utilized a U.S. verbal abilities test that provided scales in both Turkish and German.

Because both the phonological awareness and the verbal memory measures are primarily pseudoword based, it is useful to mention that the verbal test administrators were trained in the non-word pronunciation to ensure uniform verbal cues. These measures were all administered in one session in the same order: phonological awareness, pseudoword memory span, expressive vocabulary, and finally, listening comprehension. The only exceptions were the Turkish and German scales within the vocabulary measure, which were administered in alternate order from participant to participant. Test administrators recorded if the verbal responses were correct or incorrect (1 or 0) for each item[6].

[6] Although allowing the examiner to determine the students' responses as correct or incorrect could be criticized for lacking in objectivity, this is a common practice in many high quality modern psychological investigations (e.g., Proctor et al., 2005).

Phonological awareness: BAKO 1-4 - Modified (T1, T2)

A modified version of the standardized phonological awareness measure "Basiskompetenzen für Lese-Rechtschreibleistungen—BAKO" was administered as a central instrument of the verbal competencies portion of this investigation at T1 and T2 (Stock et al., 2004). The BAKO was designed to serve as a standardized test battery for primary school children in grades 1-4 for assessing phonological awareness in German-speaking countries. A substantial proportion of the BAKO test items are German non-words, although several items are in German and the instruction is also in German. To be fair to children of all language backgrounds, value was placed on language neutrality. Furthermore, the necessity of time and cost-efficiency required a phonological awareness measure that was as economic as possible. The BAKO was therefore shortened and reduced to pseudoword items wherever possible. Additionally, in order to reduce language confounds, the structures of the pseudowords were modified to be possible non-words in both Turkish and German. This required the elimination of consonant clusters not found in typical Turkish language structures.

Four of the seven BAKO subtests were included in this investigation: pseudoword segmentation, verb replacement, word remainder determination, and sound categorization. These four scales were selected based on their predominantly pseudoword composition. Since the instrument was modified for this study, item analyses were conducted on all administered items of each scale. Looking at item-scale correlations for the entire sample, no items were found that correlated with the scale under .20 and that lowered the total alpha coefficient. Thus, no items were removed from any of the four utilized scales. A copy of the exact measures can be found in Appendix A and the item-analyses can be found in Appendix B (for each group separately only).

The duration of the BAKO subscale administration was approximately 10-15 minutes in individualized form. As mentioned above, all phonological measures were administered orally. Test administrators recorded if the verbal responses were a hit or miss (1 or 0) for each item.

The *pseudoword segmentation* scale consisted of eight items that required the participant to listen to and repeat a non-word, then identify each sound (phoneme) in the word with a small plain card to aid the counting. For example, "/bareta/" would require the identification and setting down of a card for each sound (/b/ /a/ /r/ /e/ /t/ /a/). This scale had an acceptable internal consistency within the current sample as a whole: $\alpha = .79$ at T1 and $\alpha = .64$ at T2. This is substantially better than the national standardization sample of 210 second grade children tested in 2002 ($\alpha = .53$). The internal consistency in this sample can therefore be considered stronger than that demonstrated with the BAKO standardization sample.

The *vowel replacement scale* was a 12 item measure of a child's ability to verbally modify vowel sounds in a non-word. In this case, the children were asked to repeat a pseudoword and replace all /a/ sounds with an /i/sound. The participant would hear the word "/afate/" for example, and would be required to say "/ifite/". This scale proved to have an excellent average internal consistency over both times of measurement (T1 r = .92, T2 r = .93). Again, this was somewhat higher than expected since the authors of the original BAKO vowel replacement scale found an alpha coefficient of .87 for second graders.

In the *word remainder determination* subtest, participants were given the task of verbalizing a non-word with either the beginning or end phoneme missing. This seven-item scale included items such as "Repeat the word '/osarof/' without the first sound". This scale had an acceptable internal consistency of .81 at T1 and .75 at T2 for this sub-sample, substantially higher than that of the standardization sample of second graders (r = .68).

The final phonological awareness scale utilized in this investigation was the eight item *sound categorization scale.* This subtest required participants to listen to a series of four non-words and real words to determine which one began or ended with the "wrong" sound (which one did not match the others). For example, on one item, the participants heard the stimuli "/pat/-/kut/-/pit/-/pal/", and were asked to find the word with a different sound at the beginning. The scale proved to have an average internal consistency of .68 for the sub-sample and again slightly higher than the expected .61 found among children in the standardization sample.

Additionally, an *aggregate phonological awareness* scale was created with all items from the four subscales. This scale consisted of 35 items and had an excellent average internal consistency of .91 over both times of measurement for the current sample. This is comparable to the alpha of .93 found for the seven subscales in the standardization sample. For reasons of parsimony, the aggregate phonological awareness scale is used as the central measure of phonological awareness in the analyses for the present investigation. A factor analysis described in the Results section of this report provides further support for using the aggregate scale instead of each phonological awareness subscale separately. A series of analyses with the standardization sample as well as the present sample (see the Results section below) showed that the BAKO had mid-sized correlations with cognitive abilities tests, reading assessments, and teacher assessments, thus indicating its stable construct validity.

Verbal memory: Pseudoword Span Test-- Abridged (T1)

This test of verbal working memory is composed entirely of pseudowords and designed particularly for use with bilingual children (Comeau & Cormier, 2000). The Pseudoword Span Test serves as a developmental measure of verbal working memory functional in educational, research, or clinical settings with grades 1-6 without floor or ceiling effects. The instrument is intended to be administered as an individual test. A standardization with 122 Canadian French-English speaking children found the Pseudoword Span Test to be developmentally sensitive as well as sufficiently valid by way of high correlations with measures of reading performance.

For the purposes of this investigation, the instrument was shortened and modified to be language neutral for both German and Turkish speakers. This required the removal of consonant clusters. In addition, the items were written phonetically in German instead of in English to ease administration for the testers. The instrument was utilized in its abridged form as recommended by the authors for time efficiency (approximately 5 minutes).

The instrument consisted of three sets of words in total. In the first set, the participant was asked to repeat groups of one-syllable non-words (e.g., /nait/-/bim/-/teest/). In the next set the participant repeated groups of two-syllable words (e.g., /rubid/-/sigbet/-/tadding/) and in the final set, three-syllable words were presented. With the increasing number of syllables, the items increased in difficulty. Three points were possible for each item-- one for each correctly repeated non-word. The task was made more difficult by requiring the test administrators to cover their mouths with a puppet and eliminating any possible visual cues from the administrators' mouths. The test was discontinued after three consecutive completely failed items to minimize frustration among the participants.

The Pseudoword Span Test functioned well with a multicultural German school population. Since items were arranged in order of increasing difficulty, reliability was measured with the split-half method, comparing the odd and even items[7]. The 12-item measure had a Spearman-Brown split-half internal consistency coefficient of .86 for the full sample. Reliability coefficients for the Canadian sample were not available for comparison. Since this is a new instrument in the German language realm, an analysis of its factor analytic properties is provided in the Results section of this paper. A copy of the exact items are included in Appendix A and the item-analyses are in Appendix B (for each group separately).

Bilingual Verbal Abilities Test (T1 & T2)

A shortened version of the Bilingual Verbal Abilities Test (BVAT; Munoz-Sandoval, Cummins, Alvarado, Ruef, 1998) was administered in German to all children and in Turkish for the Turkish-speaking children. The BVAT scales were derived from three tests from the Woodcock-Johnson Psycho-Educational Battery-Revised and designed for use with people ages

[7] Other analyses assessing internal consistency (e.g., Cronbach's alpha) are based on the assumption that all items in the scale are equal. Since the items in the verbal memory measure increase in difficulty, only odd-even split half methods of assessing reliability are possible (Green, Salkind, & Akey, 2000).

5-90 in research, clinical assessments, or educational settings. Available in 16 different languages, the BVAT aims to provide a standardized and psychometrically sound procedure for combining verbal L1 and L2 assessment in the same instrument. Because the utilized subscales of the BVAT tapped into the participant's ability to *produce* appropriate verbal responses, they are considered measures of *expressive* vocabulary. The BVAT is a measure of cognitive academic language proficiency (CALP) assessing school-related language, as opposed to general conversational language.

In this investigation, three of the four BVAT subtests were selected based on age-appropriateness and time efficiency. The children were tested with the *picture vocabulary*, *oral vocabulary synonyms*, and *oral vocabulary antonyms* subtests in individual testing sessions. All children received the three German subtests, but Turkish-speaking children additionally received the three subtests in Turkish as well. The 30-item *picture vocabulary* subtest required the participants to orally identify a small black and white picture. Items increased gradually in difficulty since the BVAT was developed to be used with adults as well as children (e.g., from "star" to "pendulum"). Only the first 30 items in the BVAT picture vocabulary scale were selected for use after a group of expert teachers deemed the later items in the scale to be too difficult for this sample of young children. After five consecutive incorrectly named items, the test was ended to avoid discouraging the children. The *synonym* measure required participants to verbally respond to a spoken stimulus (e.g., "small") with a similar word (e.g., "tiny"). At T1, the synonym measure had a maximum of 10 items and at T2, 14 items. Administration was discontinued after four consecutively missed items. The 18-item *antonym* measure was similar but the intended response was the opposite of the stimulus (e.g., "young/old"). Test administrators were also required to discontinue after four incorrectly identified items in the antonym test. The picture vocabulary and the synonym tests were administered at both T1 and T2. At T2, additional items were added to the synonym scale and the more difficult antonym scale was included to reduce the chance of obtaining ceiling effects in case a large amount of development had taken place. The total BVAT test battery required between 8 and 12 minutes, depending on the participants' speed, concentration, and abilities.

Split-half reliability for English subtests ranged from r = .77 to r = .89 for 7 to 9-year-old children in the American standardization sample. However, neither alpha coefficients nor any other information was provided regarding the psychometric properties of the measures in German or Turkish. The authors describe in detail the meticulous eight-step translation procedure and consider it sufficient for ensuring reasonable psychometric properties in the translated instruments. Although in some languages items had to be deleted due to linguistic incompatibility, both the Turkish and German versions contained all items in the original English version. Validity of the BVAT was considered high since the English BVAT correlates highly with other English language, reading, and writing assessment measures.

Because the items in each scale are ordered to increase in difficulty, and items were thus unequal to each other in difficulty, reliability was calculated with an odd/even split-half analysis of internal consistency to ensure an equal number of easy and difficult items in each half. The split-half Spearman-Brown coefficients at T1 and T2 for German picture vocabulary subtest were .89 and .87 respectively. The Spearman-Brown coefficient was .74 for the synonym subtest at T1 and .78 at T2. For the antonym subtest, Spearman-Brown coefficient was .74 at T2 (the only time at which it was administered). Split-half reliability coefficients for the Turkish subtests were similar (picture vocabulary = .80 at T1 and .81 at T2; synonyms subtest = .65 at T1 and .79 at T2; antonyms = .85 for T2.

Item-scale correlations were analyzed and items with a correlation under .20 that lowered the total alpha coefficient of the scale were removed. This left the German picture vocabulary subtest with 27 of the original 30 items at both T1 and T2. All items in the German synonym subtest had sufficient item-scale correlations and were all kept (10 at T1 and 14 at T2). Three of the 18 items were removed from the original German antonym scale. The three Turkish BVAT scales had the same number of items at the outset, but after removing items with low item-scale correlations, 25 items remained in the Turkish picture vocabulary scale at T1 and 20 items at T2. The Turkish synonym scale retained 7 items at T1 and 9 items at T2, while the Turkish antonym measure retained 17 of its 18 original items after removal of one item with a low item-scale correlation. For longitudinal analyses of growth in any of the BVAT scales, the total scales were used to ensure valid comparisons of the measures over time. The item-analyses with an explanation of the removed items is found in Appendix B (for each group separately only).

To simplify analyses, the subscales were aggregated into a single *German expressive vocabulary* scale for each time of measurement that will be used in the central analyses of the investigation. The aggregate measure consisted of the tailored subscales in which the poorer items had been removed. The only exception to this were cases in which vocabulary development was analyzed longitudinally. Again, the retention of all items for longitudinal analyses procedure enabled longitudinal comparisons with all of the same items at each point of measurement, regardless of their item-scale correlations. The total number of items in the German aggregate vocabulary scale was 37 items at T1 and 55 items at T2. In the Turkish aggregate scale, 29 items were included at T1 and 44 items at T2.[8]

[8] The aggregate Turkish scale does not play a principal role for the research questions in this paper. Its most important purpose was to gain insight into the Turkish abilities of the bilingual children and to help identify those who did not meet the standard for basic levels of bilingualism. As described in Section 3.2.1., scores on the

Listening comprehension: Knuspels Leseverständnis: Subtest Hörverständnis (T2)

The Knuspel Reading Comprehension Test (Knuspels Leseverständnis; Marx, 1998) is a reading comprehension test for children in primary school based on a developmental model of literacy acquisition in which decoding, recoding, and listening comprehension play central roles. The test was designed primarily for use in the classroom as a diagnostic tool to detect reading difficulties. Of the four subtests, only the listening comprehension subscale was administered. This subtest was intended to measure listening comprehension for orally presented questions and directions. Two to three points were available for each item. One to two points were possible for the correct response to the content of the question and one to two points were possible for correctly carrying out instructions. For example, one item read, "Write your name in print on the line." One point was possible for each of the two tasks: one for the correct writing of the name, and the other point was contingent on the correct positioning of the name on the line (as opposed to below the line). Because the differentiation between comprehension of verbal instructions and the comprehension of verbal questions was not of interest in this investigation, but rather listening comprehension in general, the two scores were not separated as originally designated in the test handbook, but collapsed into a single listening comprehension score.

Of the 16 original Knuspel items (including two example items), 14 were selected for this study. The two items that were eliminated were more time-consuming to code as correct or incorrect and therefore not efficient for a larger scale study. However, item analyses with the full sample uncovered three items with low inter-item correlation, which were then removed from the scale. With two to three points possible for each remaining item, a total of 26 points were achievable. Neither the reliability (Cronbach's alpha = .61) nor the factor analysis (see Section 4.1.1.) for this instrument were particularly strong. Perhaps that is one reason why only retest reliability (.80 for individual testing) is available in the Knuspel handbook and no information regarding the factor analytic properties of the instrument is provided.

The greatest advantage of the Knuspel listening comprehension test was that it included children with different language backgrounds into its standardization sample. Although the manual indicates that bilingual children perform around one standard deviation under the average scores of the monolingual children, the authors indicate that the test is valid and reliable for multilingual children as well as monolingual German children. The test manual provides separate *T*-values and percent rankings for both groups. Nonetheless, this instrument is used sparingly and with caution in the following analyses due to its uncertain psychometric qualities.

aggregate Turkish scale were used to exclude those children who were more than one standard deviation away from the mean of the group of children who reported speaking Turkish or were reported to speak Turkish at home.

Finally, Table 10 provides a comprehensive overview of the instruments administered and constructs measured at each point of measurement. For review of the study's time frame and each point of measurement, see Table 1 in Section 2.3..

Table 10

Measurement Instruments at Five Times of Measurement

Time -1		**Middle of first grade**
	Cognitive abilities	Culture Free Intelligence Test
	Teacher ratings	Teacher's initial assessment
Time 0		**End of first grade**
	Word decoding	Würzburg Silent Reading Test
Time 1		**Middle of second grade**
	Teacher ratings	Teacher's assessment
	Verbal memory	Pseudoword Span Test
	Cultural capital	Measure of Home Literacy
	German/Turkish Verbal abilities	Bilingual Verbal Abilities Test (Turkish & German) ▪ *Picture Vocabulary* ▪ *Synonyms*
	Phonological awareness	BAKO 1-4 ▪ *Phoneme segmentation* ▪ *Sound categorization* ▪ *Word remainder determination* ▪ *Phoneme replacement*
	Word decoding	Würzburg Silent Reading Test
Time 2		**End of second grade**
	Verbal memory	Pseudoword Memory Span Test
	Cultural capital	Measure of Home Literacy
	German/Turkish Verbal abilities	Bilingual Verbal Abilities Test ▪ *Picture Vocabulary* ▪ *Synonyms* ▪ *Antonyms*
	Phonological awareness	BAKO 1-4 ▪ *Phoneme segmentation* ▪ *Sound categorization* ▪ *Word remainder determination* ▪ *Phoneme replacement*
	Listening comprehension	German listening comprehension
	Word decoding	Würzburg Silent Reading Test
	Reading comprehension	German reading comprehension test (ELFE)
Time 3		**Middle of third grade**
	Word decoding	Würzburg Silent Reading Test
	Reading comprehension	German reading comprehension (ELFE)

3.4. Data treatment

As is typical in longitudinal research (see Berk, 2004) and particularly in studies with children in inner-city educational settings, a certain amount of missing data was to be expected in this investigation over the five points of measurement. The measures administered exclusively for this study (phonological awareness, expressive vocabulary, verbal memory, and listening comprehension) are complete for both points of measurement. However, data collected over the five test periods within the larger BeLesen study do have some missing data. Table 11 provides a description of the number of cases available for each measure at each time of measurement[9].

Analyses of the missing values showed that there was no systematic loss of data over the five points of measurement. Using the t-tests within the Missing Values Analyses in SPSS 12.0 for the entire sample (*N* = 169) and for the two groups separately (monolingual *N* = 69, bilingual *N* = 100), analyses showed that participants who were absent for any measure at any point of time demonstrated no signs of significantly deviant performance on any of the other primary measures of interest (cognitive abilities at T-1, verbal measures at T1 and T2, decoding at T0-T3, and reading comprehension at T2 and T3). Furthermore, a separate analysis of variance for each group showed that there was no significant difference between children who were present or not present at the final point of measurement (T3) with regard to the T-1 measurement of cognitive abilities (TB: $F(1,89) = .07, p = .80$; GM: $F(1,56) = 3.18, p = .08$)[10]. In other words, there was no evidence that children of higher or lower cognitive abilities were more likely to drop out by T3. There were also no substantial differences in the number of missing cases within the two groups on any measure, with two minor exceptions: 1) There were fewer teacher surveys available for German monolingual children at T1, $\chi^2(1, N = 169) = 4.55, p = .03$, and 2) at T3, proportionally more German monolingual children were missing than Turkish-German bilingual children, $\chi^2(1, N = 169) = 6.52, p = .01$. Overall, the missing data analyses indicated that absence or attrition throughout the 24 months of investigation was fully independent of reading-related skills performance or cognitive skills. The missing data was not imputed or replaced. Instead, all analyses were conducted to allow for missing data.

All analyses, unless otherwise specified were performed with SPSS v. 12.0. The structural equation modeling program, AMOS 5 (Arbuckle, 2003), was used to test the proposed structural equation models of reading for each group in Section 4.5..

[9] For the purpose of simplification, Table 11 is displayed here in lieu of the additional provision of *N* values with each analysis for the remainder of this dissertation.

[10] A complete record of the means for the missing data analyses can be found in Appendix C: Tables C1, C2, and C3.

Table 11

Number of Participants Completing each Measure for every Point of Measurement

	N	
	TB	GM
Cognitive abilities (T-1)	90	57
Teacher assessments T-1[a]		
Language abilities	95	63
Readiness to learn	95	61
Concentration abilities	95	61
Teacher assessments T1 [a]		
Language abilities	95	60
Readiness to learn	95	59
Concentration abilities	95	59
Self-report background information		
Siblings	99	69
Parent reading behavior	100	69
Childcare experience	97	69
Verbal data T1 (Phonological awareness, verbal memory, expressive vocabulary)[b]	100	69
Verbal data T2 (Phonological awareness, expressive vocabulary, listening comprehension) [b]	100	69
Word decoding		
T0	94	63
T1	95	69
T2	95	65
T3	91	53
Reading comprehension		
T2	95	64
T3	91	53

Note. GM = German Monolingual, TB = Turkish Bilingual.

[a] In a small number of cases, single items were omitted by teachers who were unsure of the proper response for certain children.

[b] There was no missing data for the individual testing sessions used to collect data for this study (the verbal measures administered at T1 and T2). The number of cases did not vary across scales.

Since most of the measures in this study did not have standardized or population-normed scores, raw scores were used for the analyses[11]. In most analyses, gender and cognitive ability served as control variables. Unless otherwise specified, an alpha level of .05 was applied to all statistical tests. The first part of the Results section presents descriptive and preliminary analyses including tests of the instrument characteristics followed by the analyses and results for each of

[11] See Comeau and collaborators (1999) for another example of this procedure (p. 33).

the 16 hypotheses. Analyses have been divided into the same four categories of research questions as in the hypotheses section above: mean differences, differential predictive strength, patterns of development, and model fit.