Running head: PREDICTIVE VALIDITY OF PUTONGHUA CDI

Predictive Reliability of the Mandarin Chinese (Putonghua) Communicative Developmental Inventories Across Early Performance-Based Strata

by

Daniel A. Kessler

A Thesis Submitted in Partial Fulfillment of

Requirements for the Degree of Bachelor of Science

With Honors in Brain, Behavior and Cognitive Science from the

University of Michigan

2009

Advisor: Dr. Twila Tardif

Abstract

The present research examines the predictive consistency of the MacArthur Communicative Developmental Inventories (CDI) in a Mandarin Chinese (Putonghua) learning sample. Children from 8 to 20 months of age were assessed using the Mandarin Chinese version of the MacArthur Communicative Developmental Inventories on either the infant or toddler long form depending on age. Caregivers of the children filled out a checklist, rating their child's ability to comprehend and/or produce up to 680 individual words, as well as other abilities such as their gesturing and grammatical complexity. Children were assessed one year later using the CDI infant or toddler short form. Participants in the follow-up study included the bottom 10%, the middle 10% and the top 10% of children from the norming study at each age group. Predictive relationships within domains were assessed using Pearson correlations to assess whether the instrument is equally predictive for low-performing children, average-performing children, and high-performing children. Results showed overall good predictive validity, but these findings did not generalize to performance-based subgroups.

Predictive Reliability of the Mandarin Chinese (Putonghua) Communicative Developmental Inventories across Early Performance-Based Strata

The MacArthur Communicative Development Inventory (CDI) is a popular instrument for the assessment of early child language development. The inventory allows for the evaluation of children aged 8 through 30 months. Developed in the early 1990's, the CDI offers a systematic means for parent-report to be used in the assessment of language in young children (L. Fenson et al., 1993). Originally developed in English, the CDI has been translated into over forty languages and dialects. Norming work has been done or is ongoing for many of these adaptations. A research initiative spearheaded by Drs. Twila Tardif and Paul Fletcher, together with Chinese collaborators Zhixiang Zhang and Weilan Liang, adapted the CDI into the Putonghua (Mandarin Chinese) Communicative Development Inventories (PCDI) (Tardif, Fletcher, Zhang, & Liang, 2008). While the authors did extensive work with their initial sample in norming, scale reliability, content validity, and external validity, there is no published work that examines the predictive validity of the PCDI. Predictive validity is important to establish in order to lay the groundwork for longitudinal studies that use the PCDI as an early measure of language development. In English, previous studies investigating the CDI have found overall good predictive validity (Feldman et al., 2005; Luyster, Shanping Qiu, Lopez, & Lord, 2007; Reese & Read, 2000), but some authors noted variations in the strength of the predictive relationship with age (L. Fenson et al., 1993).

Language development is a rapid and at times seemingly spontaneous process. Historically psycholinguists have attempted to construct a model of the "modal child" by evaluating language development in many children in the hope of describing a typical course (L. Fenson et al., 1994). However, challenges have emerged as data has often shown great variability

in the rate of language development and onset of specific language skills (L. Fenson et al., 1994; Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991). Language development may progress in rapid advances rather than gradual, linear growth, and individual differences between children may not reflect true differences in long-term ability development but instead differences in the time that these spurts occur (as described in Scarborough, 2001). If this is the case with substantial cross-child variation in the time at which various abilities emerge, one must be very careful in controlling for age. Indeed, even the particular course of language development has been shown to vary by performance strata. For example, morphological development onset usually follows achievement of an MLU of at least 2.0; however, language-impaired children do not begin morphological development until long afterward (L. Fenson et al., 1994).

Even if issues of differential rates of development and nonstandard developmental trajectories are overcome, there are additional, statistical difficulties involved when trying to examine the predictive validity of an instrument towards the exploration of the trajectory of language development. Simple correlations between two measures separated in time, even among identical or highly related measures, can be spuriously inflated by retention of rank across time initial measures are widely spread. While an initial distribution may shift its mean upwards from an initial point to a follow-up evaluation, individual participants may be moving around within the distribution. Moreover, the spread of the distribution could be shifting dramatically. Children at the tails of the distribution may move ever further away from the central tendency, inflating correlational observations through a fan effect where the distribution becomes wider and wider with time, where the performance distribution is also tied to a distribution that reflects rate of development. Alternatively, children who were precocious early may show some regression to the mean at follow-up points as they either reach maximal development in a

domain and wait for others to catch up, as in the case of rate asynchronicity in vocabulary comprehension and word production (L. Fenson et al., 1994).

Previous work examining predictive validity of the CDI has noted these difficulties (Bornstein, Hahn, & Haynes, 2004; Feldman et al., 2005; L. Fenson et al., 1994; P. Lyytinen & H. Lyytinen, 2004; Reese & Read, 2000; Thal, Reilly, Seibert, Jeffries, & J. Fenson, 2004). Bornstein et al (2004) framed this challenge as a question of stability, operationalizing stability as the maintenance of rank order (relative to peers) from one time point to another. This approach seeks to describe global stability but may not be able to take into account regions of instability. Overall rank order might be relatively predictive of future rank order, but stability may not hold among either atypically low performing or precocious children. Concerns about local stability in specific, maximally contrastive regions of the distribution of performance (namely the center and tails) is of serious concern as many longitudinal studies of language focus on at-risk populations (e.g. Bishop, Price, Dale, & Plomin, 2003; Hadley & Holt, 2006; Thal et al., 2004) who tend to be at the lower tail of the distribution for language measures where local stability has not been examined and global stability might not hold.

Kagan (as cited in Bornstein et al., 2004) describes two kinds of stability. The first, "complete" stability, is achieved when rank order is maintained for an individual on the same measure through time, whereas "heterotypic" stability is achieved when rank order is maintained through time on different, though theoretically related, behaviors. It is important to establish both complete and heterotypic stability for language measures. Complete stability is crucial in that it allows for the modeling of growth within one specific behavior (e.g., speech production). Heterotypic stability is also important as language acquisition is not a cleanly sequenced developmental process (L. Fenson et al., 1994): children do not simply linearly increase in one

skill to ceiling (e.g., phonology or vocabulary) before moving on to improvement in other domains (e.g. grammar). As such, in some cases complete stability may be difficult to demonstrate when the time course spans a period of rapid growth wherein the behavior is wholly undeveloped at floor initially before undergoing development whereupon it is measured later (Feldman et al., 2000; Scarborough, 2001). For example, it would not make sense to explore complete stability in long-division skill between kindergarteners and fifth-graders; instead, heterotypic stability, perhaps relating counting ability to long-division, makes more sense.

Other authors have found substantial recovery where more than half of the children with significant delays at initial assessment spontaneously recovered and joined the "normal" population at or before age 3 by falling above the criteria for persistent language delays while many of the children who later met the criteria for persistent language delay did *not* come from the early delay group (Dale, Price, Bishop, & Plomin, 2003). Retrospective analysis has shown that in some cases, the children with early language delays who persist in their delay are those who demonstrate delayed gestural development and thus it is important to not only consider the words or sentences that children can comprehend or produce but their nonlinguistic communicative abilities as well (Thal, Tobias, & Morrison, 1991).

The present study is designed to maximally allow for the investigation of the issues of stability raised by the CDI. First, local stability issues aside, I seek to establish, in Mandarin Chinese (Putonghua), the overall predictive validity of the PCDI instrument at a one year follow-up. Noting the cross-domain predictive results from above (Thal et al., 1991), I will examine both "complete" and "heterotypic" stability of the instrument (as described in Bornstein et al., 2004).

Second, I examine the PCDI as a classifier of language development and its categorical predictive power in the face of questions raised by previous studies that find children with very low performance at initial assessment often spontaneously recover and that conversely many children with initially normal performance later meet criteria for language delays (Dale et al., 2003; Thal et al., 1991). This cruder, categorical, stability is critical for studies with a focus on clinical outcome (e.g., Dale et al., 2003; Thal et al., 2004).

Finally, we turn to the issue of local stability. Some of the unique features of the sampling design of the present study (discussed below) allow examination of both complete and heterotypic stability separately for different, maximally contrastive groups. By exploring the predictive validity among participants who fall at the tails and center of a distribution of language performance separately, we can make inferences regarding the consistency of stability across the distribution and make recommendations regarding the appropriate use of the PCDI in normative, average, at-risk, or precocious populations.

Moreover, little work has explored the differential predictive validity of the CDI (or PCDI) for different performance strata. Despite the strength of the predictive relationships reported previously, these relationships could be driven by the relative ordinality of the participants. In other words, the children who were at the bottom stayed at the bottom while the children at the top stayed at the top. Would these predictive relationships hold if done separately for the middle and extremes of the distribution? These are important questions to address before the PCDI can be used longitudinally, particularly if at-risk populations are included; such studies often involve children only at the lower tail of the distribution and often use repeated CDI sampling at multiple points in time (Bishop et al., 2003; Feldman et al., 2005; Luyster et al., 2005; Luyster et al., 2007). Thus, for the present study, my question is whether the PCDI is

more predictive of categorical classification or is the predictive relationship stable along all points of the relationship? More specifically, the research questions are as follows: Do the raw scores on Production increase at equal rates, or would we expect more of a fan effect? Do percentile distributions stay essentially the same? Do children stay "in category?" Within a category, do children retain their ranking?

Method

Measures

Following Fenson et al.'s structure (1993), the PCDI consists of two separate forms (Tardif et al., 2008). The PCDI Words and Gestures Form (WG), is for children between 8 and 16 months, while the PCDI Words and Sentences Form (WS) is for children between 16 and 32 months. Children aged 16 months can be assessed using either the infant or toddler forms depending on level of advancement. Additionally, there are both long and short versions of each of these forms. Both WG and WS long forms were used at entry into the study when participants were 8-20 months old. During follow-up, the production section of the WS short form and the grammar complexity section of the WS long form were administered. Details of the three forms used are below.

PCDI: Words and Gestures Long Form.

The Words and Gestures Long Form (WGLF) is designed for use with 8-16 month old infants and is comprised of two parts. Part one includes four sections (Tardif et al., 2008). The first section begins by asking three questions about the child's attention to language. The next section asks about the child's understanding of 27 common phrases. The third section asks four questions about the child's imitative and spontaneous use of words and phrases. The fourth section is a vocabulary checklist of 411 words further subdivided into 20 semantic and syntactic

categories. For each item on the list, caregivers report whether the child can understand the word and also if they can say it. In scoring, children who can say a word are presumed to also understand it. This vocabulary list is used to generate a total score for both Comprehension (total number of words can say or understand) and Production (total number of words can say).

The second part focuses on communicative gestures in five sections. Its first section includes 11 items that ask about the child's first intentional communicative gestures. The developers of the PCDI added culturally specific gestures when adapting the form from English, such as gestures for "thank you." The second section has five items that ask about games and interactive routines that children and adults play together. Together, the first two sections comprise the "Early Gestures" component of part two. The third, fourth, and fifth sections investigate the child's understanding of objects and their use and the child's imitation of adult gestures. Together, these comprise the "Later Gestures" component of part two. Altogether, the sections of part two are combined to generate a Gestures score that reflects total gestures.

In the PCDI manual, norming data is available for this form that spans 8 to 16 months and offers tables for converting raw scores to percentiles separately by gender for the Words and Gestures Long Form (Tardif et al., 2008). Internal consistency on the scales has been previously reported in the norming study (Tardif et al., 2008). For vocabulary production, Cronbach's alphas ranged from .84 to .99 for the subdivisions of the vocabulary checklist. When each of the subsections was treated as an item, overall alpha for Comprehension and Production were both .93. The Gestures portion of the instrument (part two) had an overall alpha of .95 when items were treated individually. Administration of the PCDI Words and Gestures Long Form via interview could take anywhere from 20 to 60 minutes depending on the child's level of advancement.

PCDI: Words and Sentences Long Form.

The Words and Sentences Long Form (WSLF) is designed for children aged 16 to 30 months (Tardif et al., 2008). It is composed of two parts. The first part is a list of 801 words organized into semantic and syntactic categories. For each item, caregivers are asked if the child can say the word. This part yields an overall score on vocabulary production (Production). The second part of the Words and Sentences form deals with sentences and grammar. The first section has questions about how the child refers to people, objects, and actions not present or otherwise displaced. Caregivers can respond either "often," "sometimes," or "no." The second section asks more specifically about grammatical features employed by the child, including the use of possessives, classifiers, aspect markers, and serial verbs. The third sections asks about how the child combines words into sentences and requires caregivers to provide examples of the longest sentences they have recently heard produced by the child. The fourth section is a list of 27 items that investigate the employment of increasingly grammatically complex sentences and phrases (Tardif et al., 2008). This section yields the Complexity score. Percentile norming tables are available in the manual and reliabilities have been reported in the norming study (Tardif et al., 2008). Cronbach's alpha for overall vocabulary production (part one of the form) was at .93 while the grammar complexity section had an alpha of .99. For a relatively advanced child, administration of the PCDI Words and Sentences Long Form via interview could take as long as 60 minutes.

PCDI: Words and Sentences Short Form.

The PCDI Words and Sentences Short Form (WSSF) is designed for children aged 16 to 30 months and reflects only a subset of part one of the PCDI Words and Sentences Long Form. Items were selected from the long form that were found to be of moderate difficulty across the

age ranges in the norming study, resulting in a list of 113 items on the vocabulary checklist (Tardif et al., 2008). For each item, caregivers are asked if the child says the word, yielding a score for Production (total number of words can say). Percentile norming tables are available in the manual and reliabilities for the short form were reported in the norming study. For vocabulary production, alpha was .99. Assessment takes only approximately 15 to 20 minutes using the PCDI Words and Sentences Short Form.

Participants

For the original norming study, 1,692 children ranging in age from 8 to 30 months were recruited from a list obtained from the Beijing District Health Office and assessed on either the infant or toddler long form PCDI (Tardif et al., 2008). The sixteen-month-old participants were assigned randomly to be tested on either the infant or toddler form. From this sample, a subset of 306 children from the 8 to 20 month age range representing the bottom, middle, and top scoring children was selected for follow-up. The bases for selection are discussed below in scoring. For a full breakdown of participant enrollment and form-usage by age and gender, see Table 1.

Initial and thus follow-up study exclusion criteria included medical and family considerations. Prior to the norming study, the list of potential participants was filtered to include only those who had been rated "normal" on the Denver Developmental Screening tests, had been of normal weight at birth (greater than 2,500 grams), were full-term (36 to 40 weeks gestational age at birth), had not experienced oxygen deprivation during birth, did not have congenital malformations—such as cleft palate, and had no family history of deafness (Tardif et al., 2008). When children's families were contacted via telephone, these questions were repeated, and further inquiry was made as to parent level of education, native language of caregivers, and the overall developmental progress of the child. In addition to the above medical exclusion,

participants were also excluded if they were developmentally delayed, had severe feeding problems or long hospitalization histories, or if they had caregivers who were congenitally deaf, had less than four years of primary education, or were not native speakers of Putonghua. During the face-to-face interviews the same screening criteria were asked about again to confirm the valid entry of the participant into the study. Table 2 presents mother education level of study participants against initial group designation. The same general trends in education level are observed across all three strata.

Scoring

Production, Comprehension, Gestures, and Complexity scores were calculated from the raw responses consistent with the scoring instructions published in the Chinese Communicative Development Inventories Manual as discussed above (Tardif et al., 2008).

Children's scores for T1 Comprehension, Production, and Production were converted to percentiles, separately by age, gender, and form using the fitted norming tables presented in the PCDI manual (Tardif et al., 2008). Children who had received the toddler long form at T1 had their Comprehension scores converted to percentiles using fitted tables from the PCDI manual. Scores for T2 Production from the WS short form were converted to percentiles using the norming tables from the PCDI manual (Tardif et al., 2008). Norming data was not available separately by gender and spanned age ranges from 16 to 28 months. However, norming data was not presented for 24, 26, or 27 month old infants, so percentile breakpoints were imputed linearly from neighboring age groups. Percentile scores for Grammar were not available.

From the original norming study, three groups reflecting three extreme strata (B10, M10, and H10) were culled for follow-up and inclusion in the current study based on T1 percentile in

Comprehension (if available) or Production. B10 are the bottom 10 percentile of the norming study, M10 are the middle 10 percentile, and H10 are the top ten percentile.

Furthermore, both in order to include participants whose ages exceeded the range of the norming data and to include Grammar as an output we computed age-adjusted scores on WG Comprehension, Production, and Production and WS Production for T1 and WS Production and Complexity for T2. Age-adjusted scores were the standardized residuals that resulted from a linear regression with age at T1 as the predictor. Measures from T2 were regressed against T1 as age of participants increased uniformly by 12 months from T1 to T2.

Procedure

Children's language development was assessed using the PCDI at entry into the study (T1) (ages ranged from 8 to 20 months) and again one year later at time two (T2) (with ages then ranging from 20 to 32 months).

Unlike the original CDI norming study which employed mailed forms filled out at home by parents, the present study used an interview-style structure to encourage thorough participation by the child's primary caregiver as designated by the family (Tardif et al., 2008). All interviews with the children and their caregivers were conducted by native Mandarin- (Putonghua-) speaking research assistants (trained in developmental psychology or pediatrics) and took place individually in the children's own homes, with occasional testing sessions taking place at child health care settings, preschools, and elementary schools (Tardif, 2003). Each session with the children and their families lasted from 30 to 60 minutes. Children were given frequent breaks between tasks. Children were tested in two sessions, no more than two weeks apart, at both T1 and T2.

Results

Descriptive Results

Mean raw scores on measures are presented graphically against age by group designation in Figures 1 through 5. Participant numbers, mean age, mean raw measure score and standard deviation are all reported in Table 3. Average raw scores and percentiles for each measure of interest are presented by gender in Table 4. On every raw measure, females outperform males. However, this is likely reflective of differences in age distributions (presented in Table 1), as the comparisons between mean percentiles is less striking with the notable exception of T2 WS Production whereby girls dramatically outperform boys.

Overall Stability from T1 toT2

The overall predictive validity of the T1 to T2 PCDI measures was evaluated in a variety of ways to maximally leverage the unique sampling design while overcoming some of its potential confounds and limitations. Because the study is both cross-sectional and longitudinal with the age cross-sections cutting through a period of very rapid growth (see Figures 1 through 5 for a presentation of raw scores against age by strata), raw scores must be adjusted for age. However, during early language acquisition, nonlinearity in the rate of language growth and variability in developmental trajectories makes it difficult to find a standard measure to examine longitudinal predictive relationships (see M10 trace in Figure 5 for an example of nonlinear growth). Additionally, given the age range, two different instruments are required to prevent younger participants from performing at floor levels of the instrument and older participants from reaching ceiling which would otherwise result in very low variability making statistical correlation theoretically difficult.

First, percentile scores (described above) were used as one means of controlling for age. Three measures from the PCDI WGLF administered at T1—Comprehension, Production, and Production—were explored as predictors for Production on WSSF at T2. Predictive relationships between these measures were explored using Pearson correlations. All T1 measures were significantly related to T2 Production percentile (see first three columns of Table 5).

The previously discussed analysis unfortunately did not include participants who were assessed on the WSLF at T1. This narrowed the sample to only include children who were 8 to 16 months old at T1 (and 20 to 28 months old at T2). To include these older participants in investigations of predictive validity, two cross-form variables were computed. A cross-form Production measure (CF Production) that reflected the percentile of Production from either the WGLF or WSLF at T1 and a cross-form "primary linguistic measure" (CF PrimLing) variable were calculated that reflected the percentiles that had been used for initial group designation. CF PrimLing was the percentile score from T1 on Comprehension for participants who had received the WGLF and the percentile score on Production for participants who had received the WSLF. Pearson correlations were used to explore relationships between the two cross-form scores and Production at T2 (see Table 5 for results). T1 CF Production and T1 CF PrimLing were both significantly related to T2 Production percentile.

An alternate form of age-control was performed through calculation of standardized residuals from a linear regression of all linguistic measures against age at T1 (discussed above). To explore predictive relationships among these age-controlled scores, we again turned to Pearson correlations between T2 Production and Grammar and T1 WGLF Comprehension, Production, and Gestures. T1 Comprehension and T1 Gestures were both significantly positively correlated with T2 Production, but notably, T1 Production was *not* related to T2 Production

when using these age-adjusted scores. For T2 Grammar, only T1 Gestures was significantly correlated, although T1 Comprehension and Production both approached significance.

Again, in an effort to maximize data utilization through the inclusion of participants tested on the WS form at T1, two cross-form measures were generated that were identical to those discussed above except that they came from the standardized residuals rather than percentile scores. Both were significantly related to T2 Production and Complexity age-adjusted scores (see Table 5 for full results).

Group Stability over Time

Stability of categorical group designation over time was also of interest, so I sought to describe development drift between categorizations. First, the T2 measure percentile scores were explored visually to see if they lent themselves to the same trimodal distribution that the T1 measures show as a result of a deliberate sampling approach (see Figures 9 and 10 for visualization of percentiles). Participants received a group designation for T2 Production based on a tertile split of their percentile score. Cutpoints for Production were at the 35^{th} and 70^{th} percentiles. A chi-square analysis overwhelming rejected the independence of the two categorization strategies χ^2 (4, N = 232) = 27.1, p<.01. While many participants changed group designation from T1 to T2, strong predictive associations emerged (see Table 6). Of participants who began in B10, more than expected (if the two categorization strategies were independent) retained their low classification against T2 Production. Slightly more B10 participants than expected under independence moved into the middle category, but far fewer made it into the high classification. Somewhat surprisingly, a fair number of the M10 participants dropped into the low group at T2. The H10 group was especially stable with the majority of participants retaining their high classification with only a few dropping to the lowest group. It appears that while

children do move around from tertile to tertile, there is a good degree of stability, especially among precocious children.

Within-Group Stability from T1 to T2

Each performance strata still contained a full age-range, so it continued to be necessary to control for age in all analyses. The analyses performed to explore within-group stability from T1 to T2 are identical as those performed for overall stability except that all analyses were performed separately by initial group designation.

Table 7 presents the results of Pearson correlations of percentile scores from T1 (Comprehension, Production, Gestures, CF Production, and CF PrimLing) against T2 WSSF Production percentile. While the only significant relationship that emerged was T1 WG Production and T1 WS Production, it should be noted that for B10 and H10 most of the correlation coefficients are directionally consistent with those observed in the examination of overall stability and in many cases approach the same magnitude. The lack of statistical significance may be attributed to the loss of power when splitting the sample into three independent groups. M10 does not show any correlations of a substantial magnitude with the exception of a curious negative relationship between T1 WG Comprehension and T2 WS Production. Highlights of these analyses are also presented graphically as scatter plots in Figures 6 through 8.

Table 8 displays the results of the correlations run with the age-adjusted standardized residuals discussed in Scoring. Now that T2 WSLF Grammar can be included as an output measure, significant relationships begin to emerge in all three groups. T1 Gestures is a strong predictor for Production at T2 for all three groups (though only statistically significant among B10 and H10). T2 Grammar is strongly predicted by all three T1 WG measures among M10 and

moderately predicted among H10. The cross-form measures are not especially predictive except for T1 CF Production strongly predicting age-adjusted T2 Grammar.

Discussion

The present study explored the predictive validity, both within and cross-domain of the recently developed Putonghua Communicative Developmental Inventories. It evaluated the stability of continuous, age-adjusted measures and the heterotypic stability of categorical classification over a one year interval. It also investigated the stability of predictive relationships across performance strata.

When using the entire sample, we found strong stability, both complete and heterotypic. Complete stability could only be assessed for production, but the overall correlation was strong with r=.257. Other studies have reported higher coefficients of correlation in studies of predictive validity of the CDI. Reese & Reed (2000) reported coefficients ranging from 0.61 to 0.75, but this stability was assessed across only six months. The original CDI manual presents data on predictive validity that crosses forms in a manner similar to the present study but also with only a six month window between testing (L. Fenson et al., 1993). The CDI's study found overall correlations from 16-24 months to 22-30 months at r=.71. Interestingly, when they broke the correlations down by age, correlations at the lower end of their age range was substantially lower with r=.53 for the 16-month-olds. One study did explore predictive validity with a one year follow-up from the CDI WSLF to the CDI-III (Feldman et al., 2005). In predicting T2 vocabulary from T1 vocabulary production, the authors note a significant correlation of r=.58.Participants in this study were drawn from an apparently healthy, normative sample. This correlation is closer in magnitude to the one reported in the present study, and this may be the result of the similar method of using a one year follow-up. However, these authors looked at

significantly older children, and as Fenson et al. (1993) noted, predictive validity is much weaker for younger than older children. While the present study's measures of stability did not reach the magnitude reported in other studies of predictive validity for the CDI, it was still significant and in the same direction. The relative weakness of our statistic may reflect first the longer interval between testing, as the present study had a full year between testing sessions as opposed to the six months in the above-reviewed studies. It may also reflect the special difficulties imposed by changing form midstream (as Reese & Reed did not). None of our forms are repeatedly exactly at T2; even the T1 WS Long Form is replaced by the Short Form at T2. It may also be a result of our inclusion of participants at the outer age ranges for which the diagnostic tool is intended whereas some of the other studies tended to sample towards the sample of the possible age range for the instrument. Indeed, Fenson et al (1993), in their discussion of predictive validity, note that stability seems low when initial measurement occurs earlier than 11 months. Finally, it may also reflect our unusual sampling procedure wherein we extracted only the most centrally tending participants and the extremes at either end of the distribution. The growth profiles for these groups alone, and taken as a whole, may not be perfectly reflective of a "typical" developmental trajectory.

The results of our exploration of group stability were mixed. While there appears to be considerably upward mobility for initially low performing children into the middle range at T2, children who started in M10 actually have outcomes similar to the children in B10. It may be that children at the upper end of the B10 group experienced only modest delays in language development that may be quickly remedied. Our H10 group was relatively stable, with the majority of children retaining their high ranking. Several did slip down into the middle range, but very few fell all the way down to the low range. One other study found similar results when

trying to predict clinical outcome based on a binary classification of initial CDI performance; namely, children who begin at the lower ranges have a moderate chance of remaining there, but initial "normal" classification status is not highly protective against later problems (Feldman et al., 2005). This echoes our observation that the T2 distribution patterns appear similar for both B10 and M10. However, some of the poor predictability may be explained by the inclusion of young children under age 11 months. This is an age range reported in the CDI Manual (L. Fenson et al., 1993) to have low predictive validity. For such children it may be more useful to include more behaviors associated with linguistic development, such as gestures (Thal et al., 1991), into the predictive model as many of the children may others perform at floor for the instrument being employed.

The strong stability observed between T1 and T2 for the overall sample cannot be cleanly generalized to the subgroups identified based on performance at T1. The overall pattern is somewhat mirrored by the B10 and H10 groups in heterotypic and complete stability for predicting production percentiles, but the M10 group does not seem to show any particular stability from T1 to T2. Taken as a whole, our findings suggest that further work needs to be done with the PCDI, and perhaps other versions of the CDI, in order to explore the stability of predictive relationships along continuum of initial performance. While there is a striking difference in predictive relationships observed in the B10 and M10 groups, when examined categorically, B10 and M10 appear to assort themselves similarly. There are unique phenomena that impact predictive validity occurring at these mini-sections of the distribution of normal performance.

References

Bishop, D. V. M., Price, T. S., Dale, P. S., & Plomin, R. (2003). Outcomes of early language delay: II. Etiology of transient and persistent language difficulties. *Journal of Speech, Language, and Hearing Research*, *46*, 561.

Bornstein, M. H., Hahn, C., & Haynes, O. M. (2004). Specific and general language performance across early childhood: Stability and gender considerations. *First Language*, *24*, 267-304. doi: 10.1177/0142723704045681.

Dale, P. S., Price, T. S., Bishop, D. V. M., & Plomin, R. (2003). Outcomes of Early Language Delay: I. Predicting Persistent and Transient Language Difficulties at 3 and 4 Years. *Journal of Speech, Language & Hearing Research*, *46*, 544-560. doi: Article.

Feldman, H. M., Dale, P. S., Campbell, T. F., Colborn, D. K., Kurs-Lasky, M., Rockette, H. E., et al. (2005). Concurrent and Predictive Validity of Parent Reports of Child Language at Ages 2 and 3 Years. *Child Development*, *76*, 856-868. doi: 10.1111/j.1467-8624.2005.00882.x.

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E., & Paradise, J. L. (2000). Measurement Properties of the MacArthur Communicative Development Inventories at Ages One and Two Years. *Child Development*, *71*, 310-322. doi: 10.1111/1467-8624.00146.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., et al. (1994). Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development*, *59* , i-185.

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D. J., Bates, E., Hartung, J., et al. (1993). *MacArthur Communicative Development Inventories: User's Guide and Technical Manual*. Singular Publishing Group San Diego, CA.

Hadley, P. A., & Holt, J. K. (2006). Individual differences in the onset of tense marking: A growth-curve analysis. *Journal of Speech, Language and Hearing Research*, *49*(5), 984.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, *27*, 236-248. doi: 10.1037/0012-1649.27.2.236.

Luyster, R., Richler, J., Risi, S., Hsu, W., Dawson, G., Bernier, R., et al. (2005). Early Regression in Social Communication in Autism Spectrum Disorders: A CPEA - Study. *Developmental Neuropsychology*, *27*, 311. doi: 10.1207/s15326942dn2703_2.

Luyster, R., Shanping Qiu, Lopez, K., & Lord, C. (2007). Predicting Outcomes of Children Referred for Autism Using the MacArthur--Bates Communicative Development Inventory. *Journal of Speech, Language & Hearing Research*, *50*(3), 667-681. doi: 10.1044/1092-4388(2007/047).

Lyytinen, P., & Lyytinen, H. (2004). Growth and predictive relations of vocabulary and inflectional morphology in children with and without familial risk for dyslexia. *Applied Psycholinguistics*, *25*, 397-411.

Reese, E., & Read, S. (2000). Predictive Validity of the New Zealand MacArthur Communicative Development Inventory: Words and Sentences. *Journal of Child Language*, *27*, 255-266. doi: 10.1017/S0305000900004098.

Scarborough, H. S. (2001). Connecting early language and literacy to later reading (dis) abilities: Evidence, theory, and practice. *Handbook of early literacy research*, *1*, 97–110.

Tardif, T. (2003). From First Words to Fluency in Mandarin: Narrative Version. Unpublished Manuscript.

Tardif, T., Fletcher, P., Zhang, Z. X., & Liang, W. L. (2008). *Chinese Communicative Development Inventories (Putonghua and Cantonese versions): User's Guide and Manual*. Beijing, China: Peking University Medical press.

Thal, D., Tobias, S., & Morrison, D. (1991). Language and Gesture in Late Talkers: A 1-Year Follow-up. *J Speech Hear Res*, *34*(3), 604-612.

Thal, D., Reilly, J., Seibert, L., Jeffries, R., & Fenson, J. (2004). Language development in children at risk for language impairment: Cross-population comparisons. *Brain and Language*, *88*(2), 167-179. doi: 10.1016/S0093-934X(03)00096-8.

Author Note

Daniel A. Kessler, Department of Psychology, University of Michigan, Ann Arbor.

The author would like to thank Dr. Twila Tardif for her hard work in laying the foundation for the analyses that I have presented here. Without her tireless efforts in the development and translation of the PCDI, extensive work writing up the norming data, and exhaustive labors in collecting initial and longitudinal follow-up data I would not have been able to write this paper. Her mentorship and guidance have steered my academic career for the past three years and I look forward to our continued working relationship and eventual colleagueship. She has spurred me on to ambitious endeavors, identified and worked with me on my weaknesses, and inspired me with her dedication, intellectual curiosity, and boundless energy. Her direct oversight of this thesis has been substantial as I undertook a topic about which I had little background in prior to this academic year, but her passion for her work and deep involvement in her field is contagious. I would also like to thank Dr. Tardif's collaborators and students who helped to develop and norm the PCDI and worked on the follow-up study.

One research assistant from Dr. Tardif's lab, Annette Leung, worked tireless long-hours with me, helping me to score data, format graphs, and take care of so many details, and I'd like to thank her here.

I have been blown away by the support that I have received from my family, friends, and professors during my thesis writing endeavors. Their understanding and patience with me as I became increasingly cranky and exhausted surprised me and has showed me how to treat otherwise difficult people (myself). I would like to especially thank my mother, whose passion for language, teaching, learning, and discovery I carry with me always, for her direct

involvement with some of the final editing of my thesis, and my father, whose gift and proclivity for quantitative analysis I rely on everyday.

Finally, I would like to take a moment to give general appreciation for the campus community. I cannot imagine a more intellectually engaging, diverting, and exciting place to work on scholarly endeavors. From the hockey team that provided me hours of weekend entertainment (and sometimes frustration) to the statistical workshops offered, the University of Michigan is a total package.

Correspondence concerning this article should be sent to Dr. Twila Tardif, University of Michigan, Center for Human Growth and Development, 300 North Ingalls Building, Ann Arbor MI, 48109.

Table 1

Participant Enrollment and Form Usage by Age and Gender

Age in Months	Males	Females	Total
	Words and Gestures		
8	14	8	22
9	17	5	22
10	10	14	24
11	13	11	24
12	10	13	23
13	14	8	22
14	13	10	23
15	13	12	25
16	13	14	27
Total	117	95	212
	Words and Sentences		
16	14	6	20
17	10	9	19
18	10	10	20
19	5	12	17
20	16	2	18
Total	55	39	94

Table 2

Mother Education Level by Initial Group Designation

Education Level	Bottom 10%	Middle 10%	High 10 %	Total
	Number of Participants			
< 3 Years	1	0	0	1
4-6 Years	1	1	0	2
7-9 Years	17	8	5	30
10-12 Years	40	38	43	121
Technical College	27	37	24	88
University	17	14	23	54
Post-University	2	4	4	10
Total	105	102	99	306

Table 3

Summary of Raw Measures

Measure	N	Mean Age	Mean Score	SD
T1 Words and Gestures				
Comprehension	212	12.12	209.60	7.40
Production	212	12.12	23.03	.96
Gestures	212	12.12	31.59	3.58
T1 Words and Sentences				
Production	94	17.94	247.03	24.87
T2 Words and Sentences				
Production	306	25.91	88.30	27.37
Grammar Complexity	285	26.09	55.95	19.51

Note. Mean age, mean score, and standard deviations have been rounded to the nearest hundredth.

Table 4

Summary of Measures (Raw and Percentiles) by Gender

Measure	Mean Male Score (SD)	Mean Female Score (SD)	Mean Male Percentile (SD)	Mean Female Percentile (SD)
		Words and Gestures		
Comprehension	207 (104)	213 (112)	52 (32)	50 (35)
Production	18 (40)	29 (64)	58 (24)	54 (25)
Gestures	30 (14)	34 (13)	47 (31)	48 (31)
		Words and Sentences		
Production	84 (30)	93 (23)	46 (31)	58 (27)
Grammar Complexity	53 (21)	60 (17)	n/a	n/a

Note. Scores, percentiles, and standard deviations have been rounded to the nearest whole.

Table 5

Correlations from T1 to T2 for Percentile and Age-Adjusted Scores

Measure 1	Measure 2	Percentiles	Age-Adjusted Standardized Residuals
		Pearson R	
T1 WG Comprehension	T2 WS Production	0.315**	.252**
T1 WG Production	T2 WS Production	0.257**	0.118
T1 WG Gestures	T2 WS Production	0.320**	0.323**
T1 WG Comprehension	T2 WS Grammar Complexity	n/a	0.120
T1 WG Production	T2 WS Grammar Complexity	n/a	0.130
T1 WG Gestures	T2 WS Grammar Complexity	n/a	0.211**
T1 CF Production	T2 WS Production	0.265**	0.149**
T1 CF PrimLing	T2 WS Production	0.317**	0.249**
T1 CF Production	T2 WS Grammar Complexity	n/a	0.204**
T1 CF PrimLing	T2 WS Grammar Complexity	n/a	0.201**

Table 6

Initial Group Designations versus T2 Production Percentile Tertiles

Initial Designation	T2 Production Tertile			Total
	Low	Middle	High	
	Observed (Expected)			
B10	35 (28)	30 (28)	15 (24)	80
M10	34 (28)	28 (28)	18 (24)	80
H10	11 (25)	24 (25)	37 (22)	72
Total	80	82	70	232

Table 7

Correlations between T1 and T2 using Percentile Scores by Group

Measure 1	Measure 2	B10	M10	H10
		Pearson R		
T1 WG Comprehension	T2 WS Production	0.169	-0.124	0.064
T1 WG Production	T2 WS Production	0.183	0.048	0.241*
T1 WG Gestures	T2 WS Production	0.175	0.080	0.151
T1 CF Production	T2 WS Production	0.177	0.059	0.210
T1 CF PrimLing	T2 WS Production	0.133	-0.084	0.058

Table 8

Correlations between T1 and T2 using Age-Adjusted Standardized Residuals by Group

Measure 1	Measure 2	B10	M10	H10
		Pearson R		
T1 WG Comprehension	T2 WS Production	0.042	0.018	0.147
T1 WG Production	T2 WS Production	-0.108	0.022	0.097
T1 WG Gestures	T2 WS Production	0.243*	0.173	0.241*
T1 WG Comprehension	T2 WS Grammar Complexity	-0.139	0.263*	0.089
T1 WG Production	T2 WS Grammar Complexity	-0.035	0.105	0.131
T1 WG Gestures	T2 WS Grammar Complexity	0.157	0.231	0.143
T1 CF Production	T2 WS Production	0.040	-0.005	0.026
T1 CF PrimLing	T2 WS Production	0.070	-0.038	0.078
T1 CF Production	T2 WS Grammar Complexity	0.320**	0.062	0.079
T1 CF PrimLing	T2 WS Grammar Complexity	-0.006	0.073	0.041

Figure Captions

Figure 1. T1 Words and Gestures Long Form: Comprehension by Age and Group.

Figure 2. T1 Words and Gestures Long Form: Production by Age and Group.

Figure 3. T1 Words and Gestures Long Form: Gestures by Age and Group.

Figure 4. T1 Words and Sentences Long Form : Production by Age and Group.

Figure 5. T1 Words and Sentences: Grammar Complexity by Age and Group.

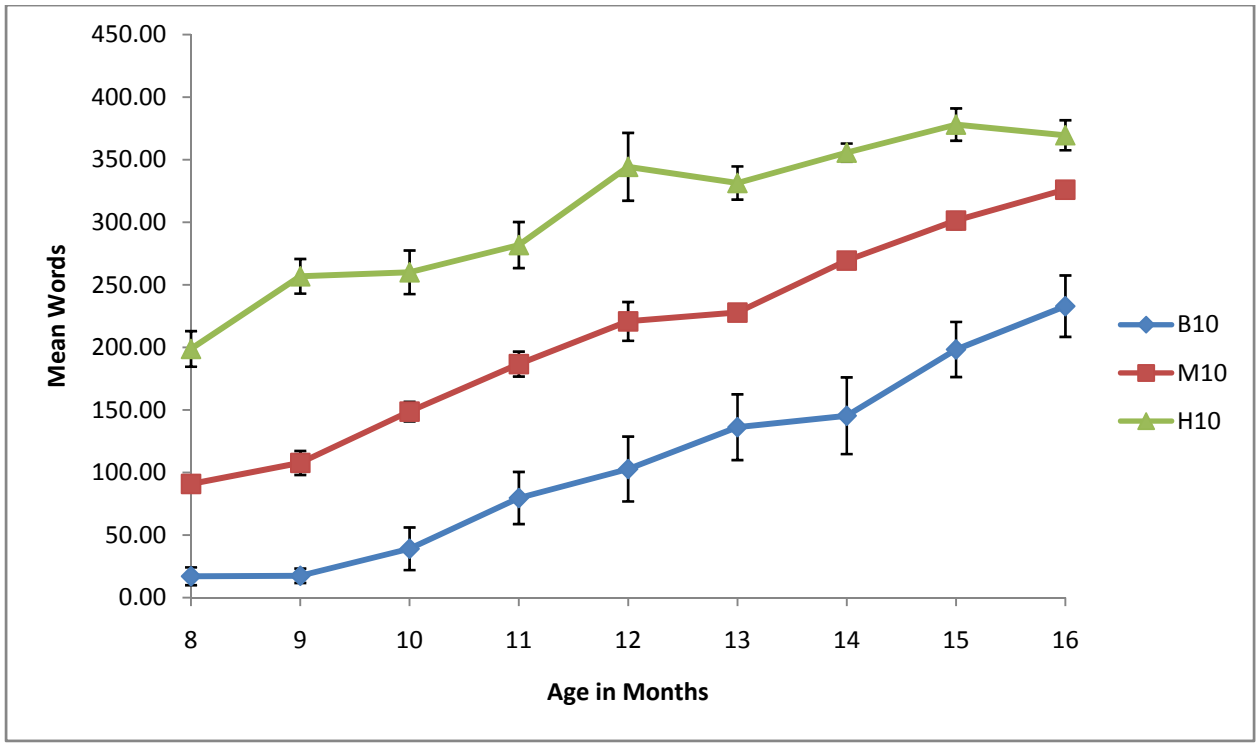
Figure 6. T1 WG Comprehension Against T2 WS Production by Group.

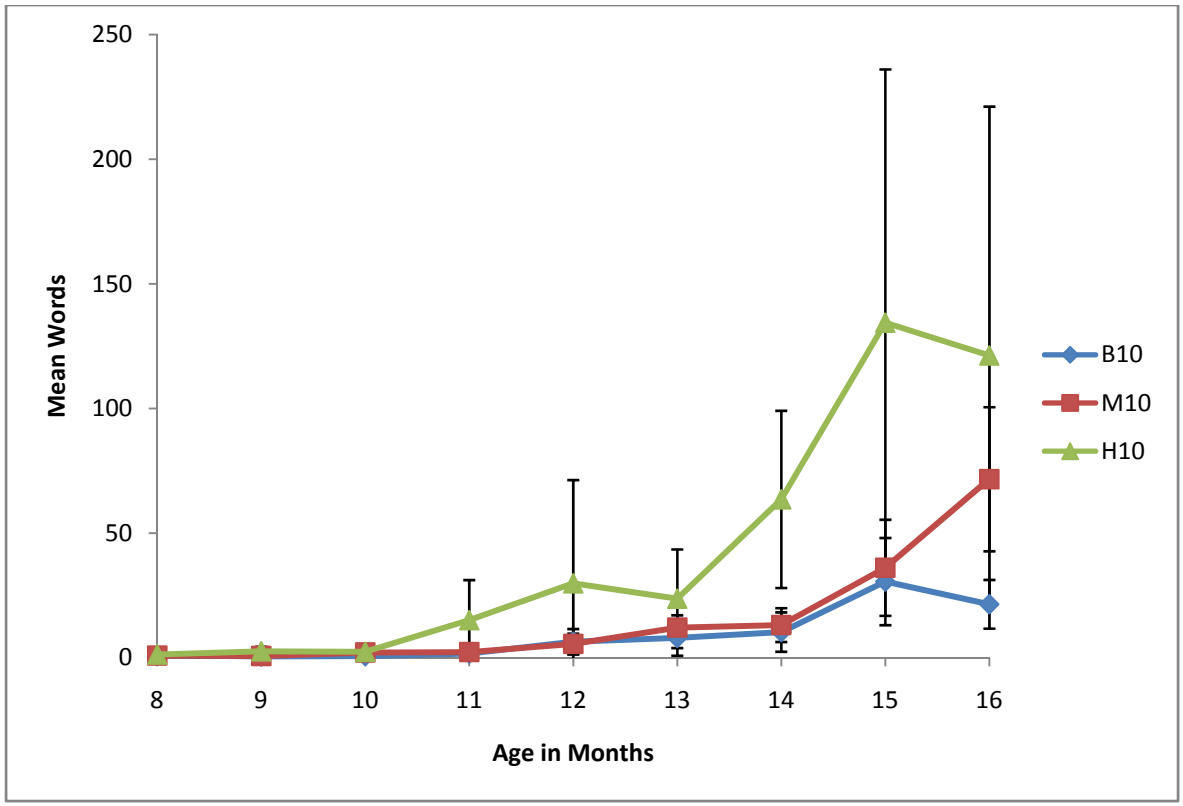
Figure 7. T1 WG Comprehension Against T2 WS Production by Group.

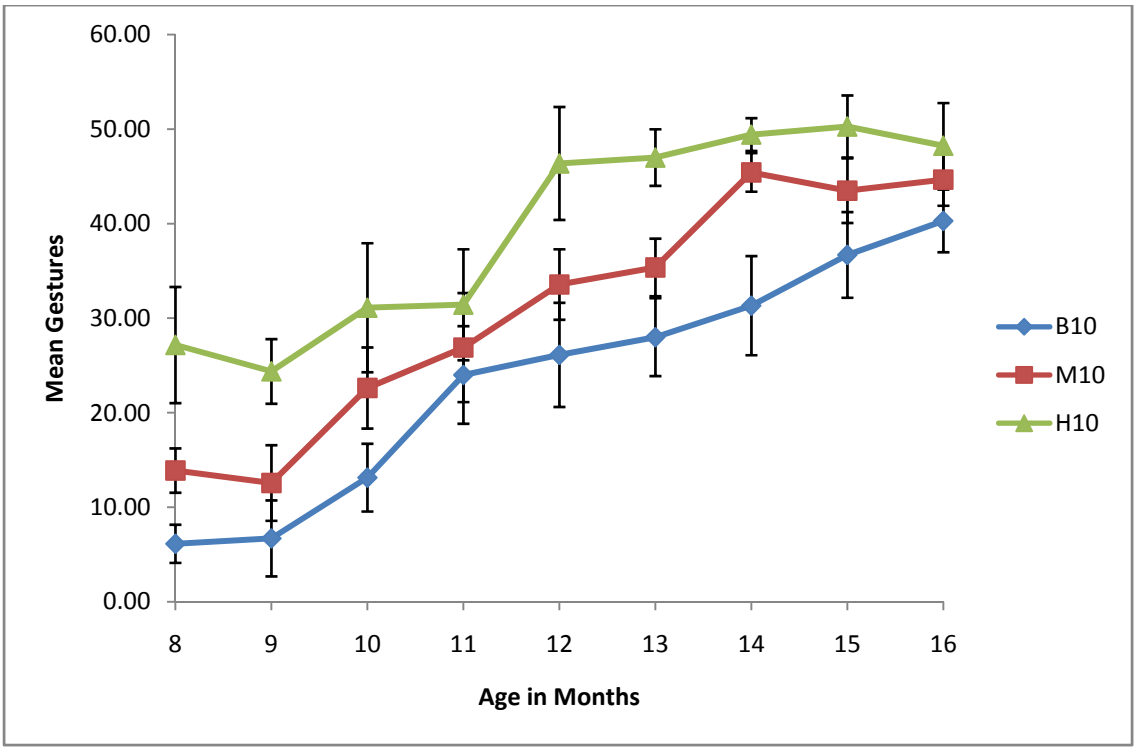
Figure 8. T1 WG Gestures Against T2 WS Production by Group.

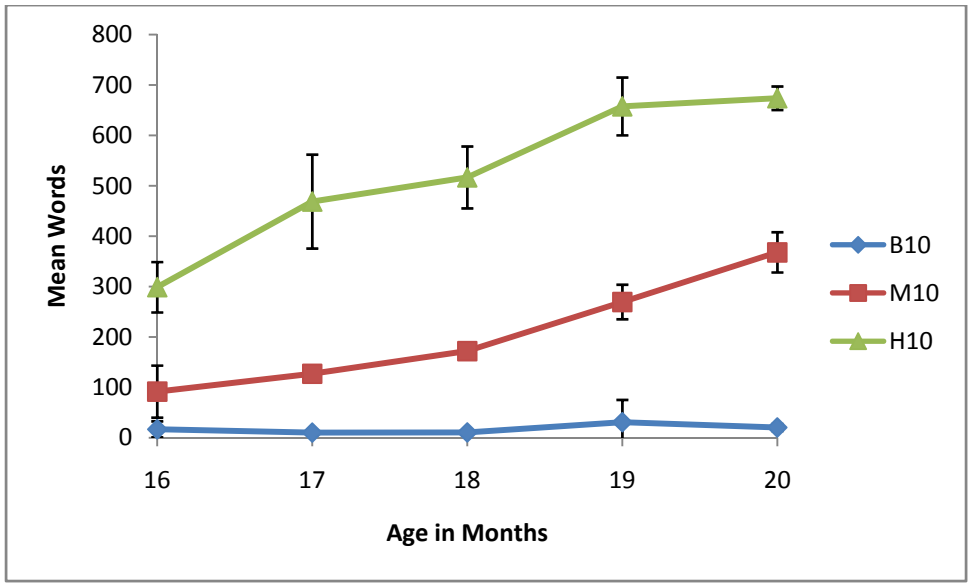
Figure 9. Evidence of a trimodal distribution (based on sampling selection) for T1 WGLF Comprehension.

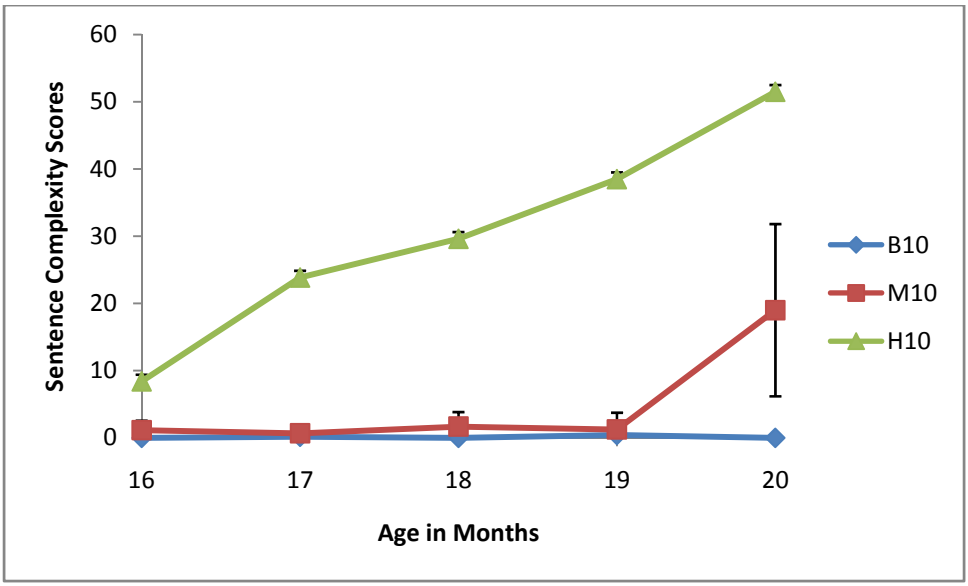
Figure 10. Substantially more normally distributed performance for T1 WSSF Production.

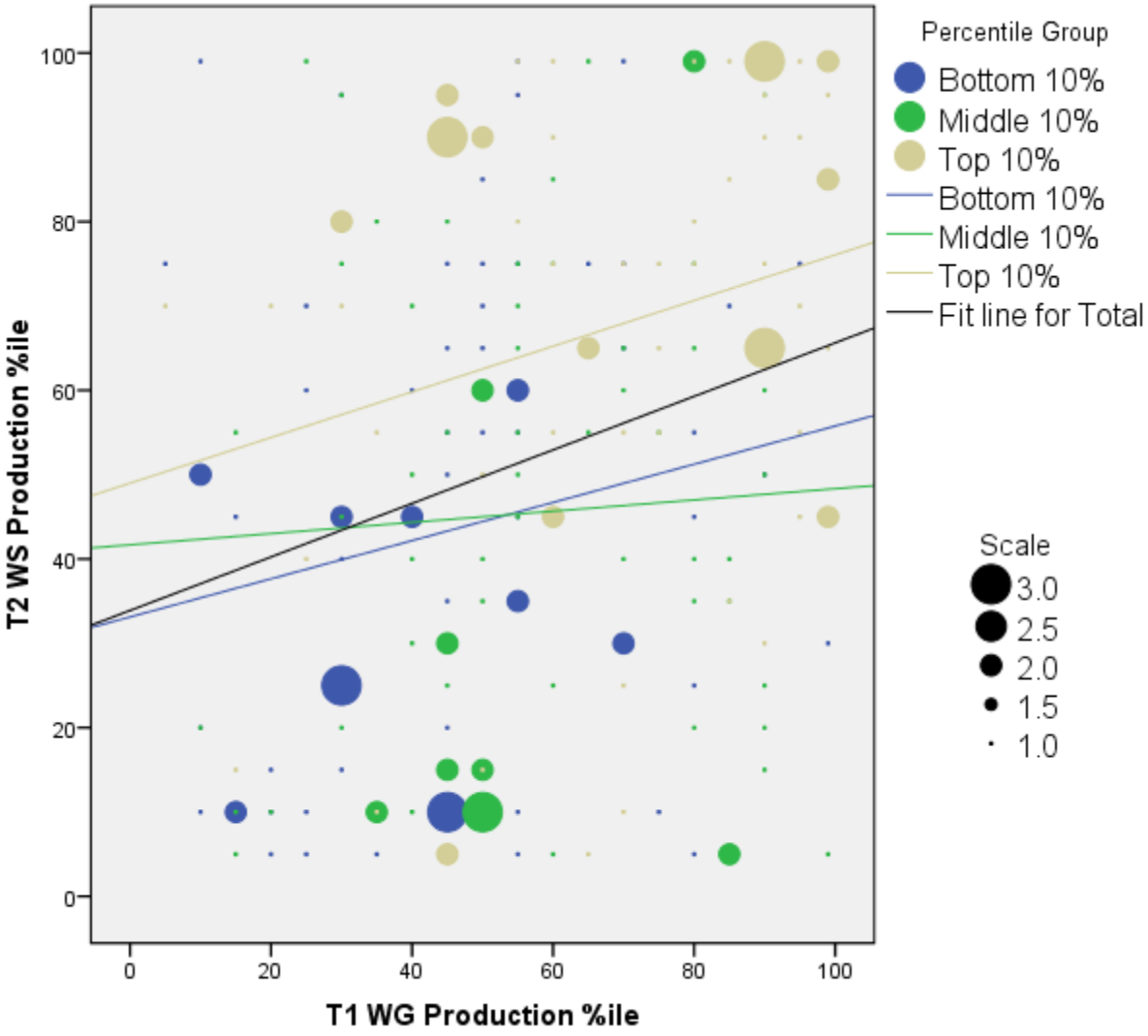

450.00
400.00
350.00
300.00
250.00
200.00
150.00
100.00
50.00
0.00
Mean Words
8
9
10
11
12
13
14
15
16
Age in Months
B10
M10
H10

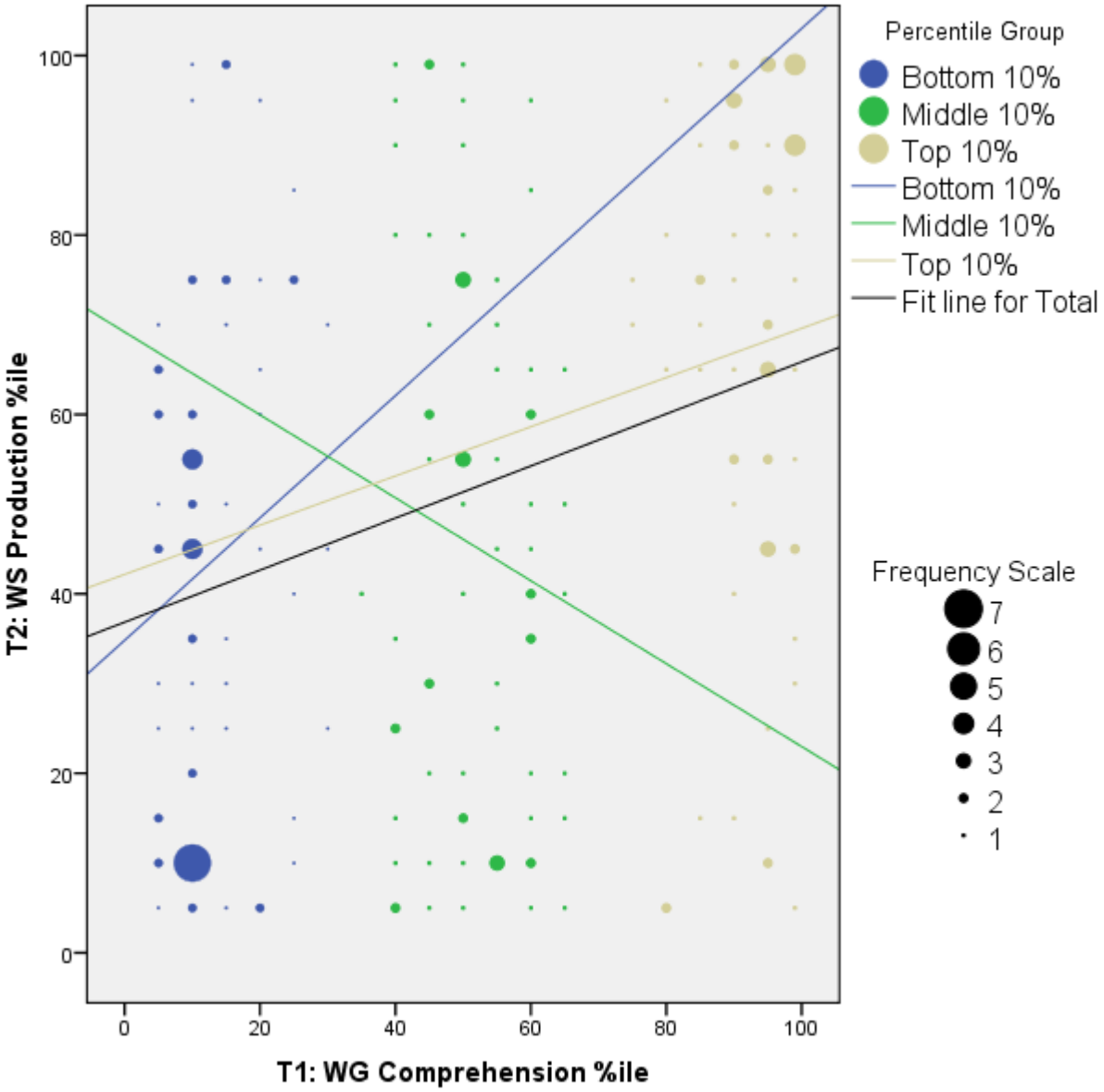

250
200
150
100
50
0
Mean Words
8
9
10
11
12
13
14
15
16
Age in Months
B10
M10
H10

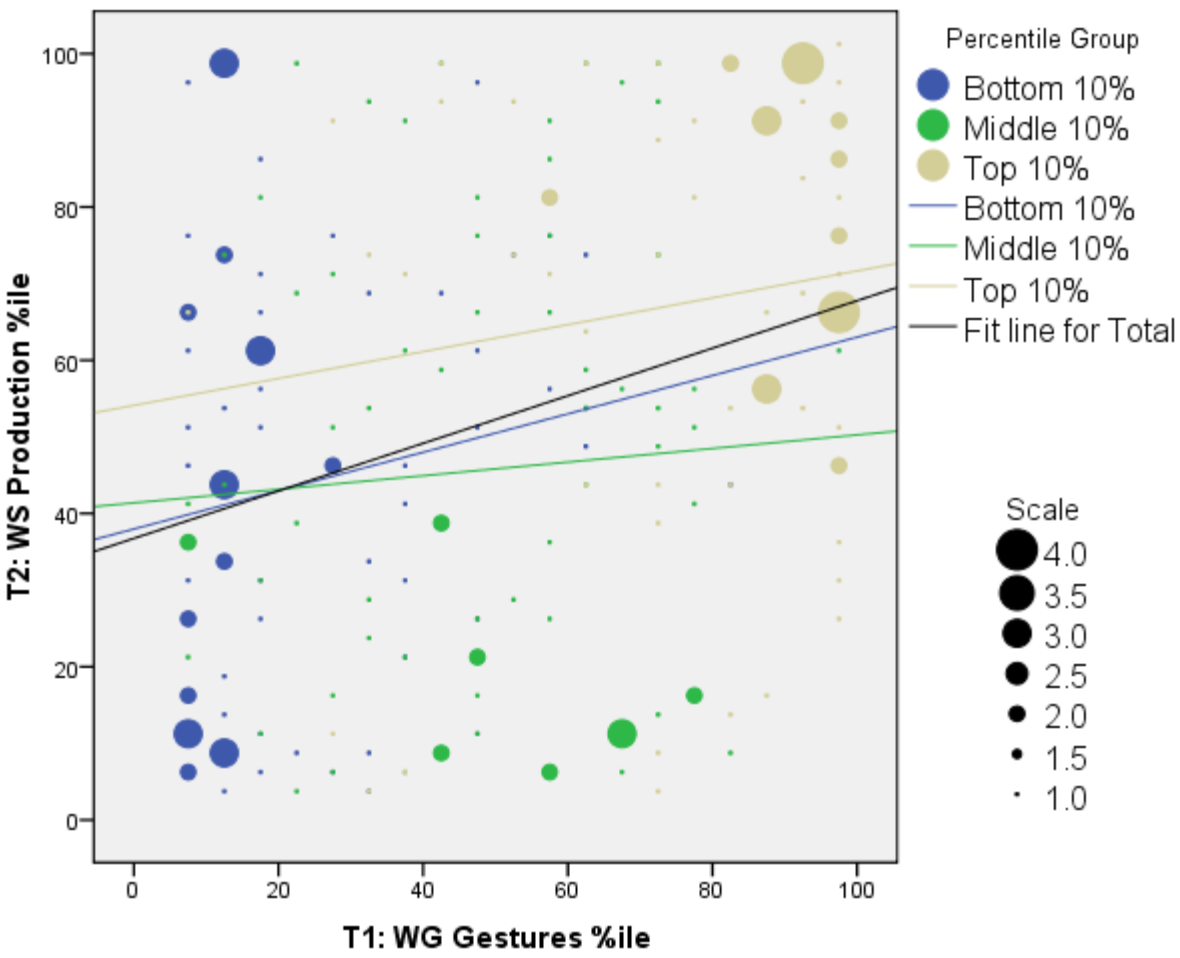

60.00
50.00
40.00
30.00
20.00
10.00
0.00
Mean Gestures
8
9
10
11
12
13
14
15
16
Age in Months
B10
M10
H10

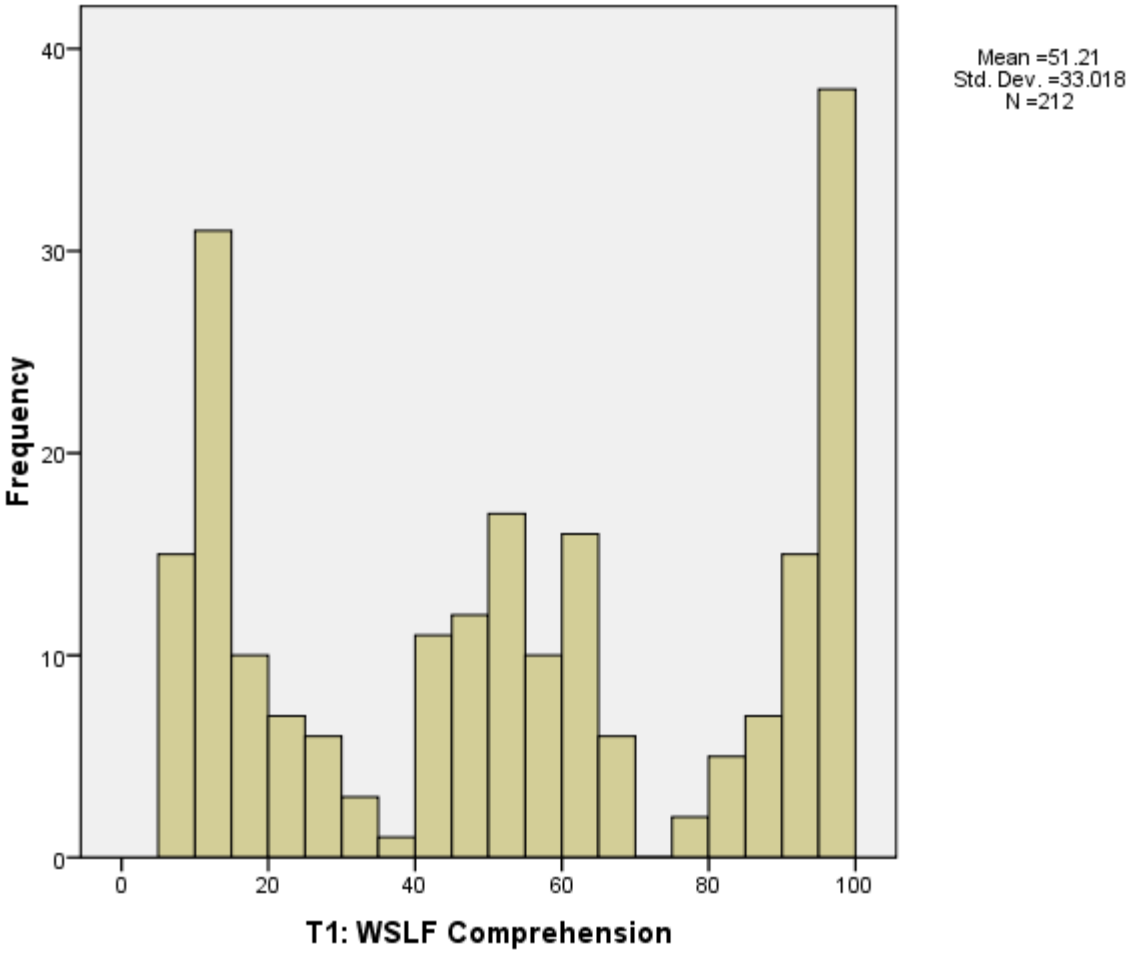

800
700
600
500
400
300
200
100
0
Mean Words
16
17
18
19
20
Age in Months
B10
M10
H10

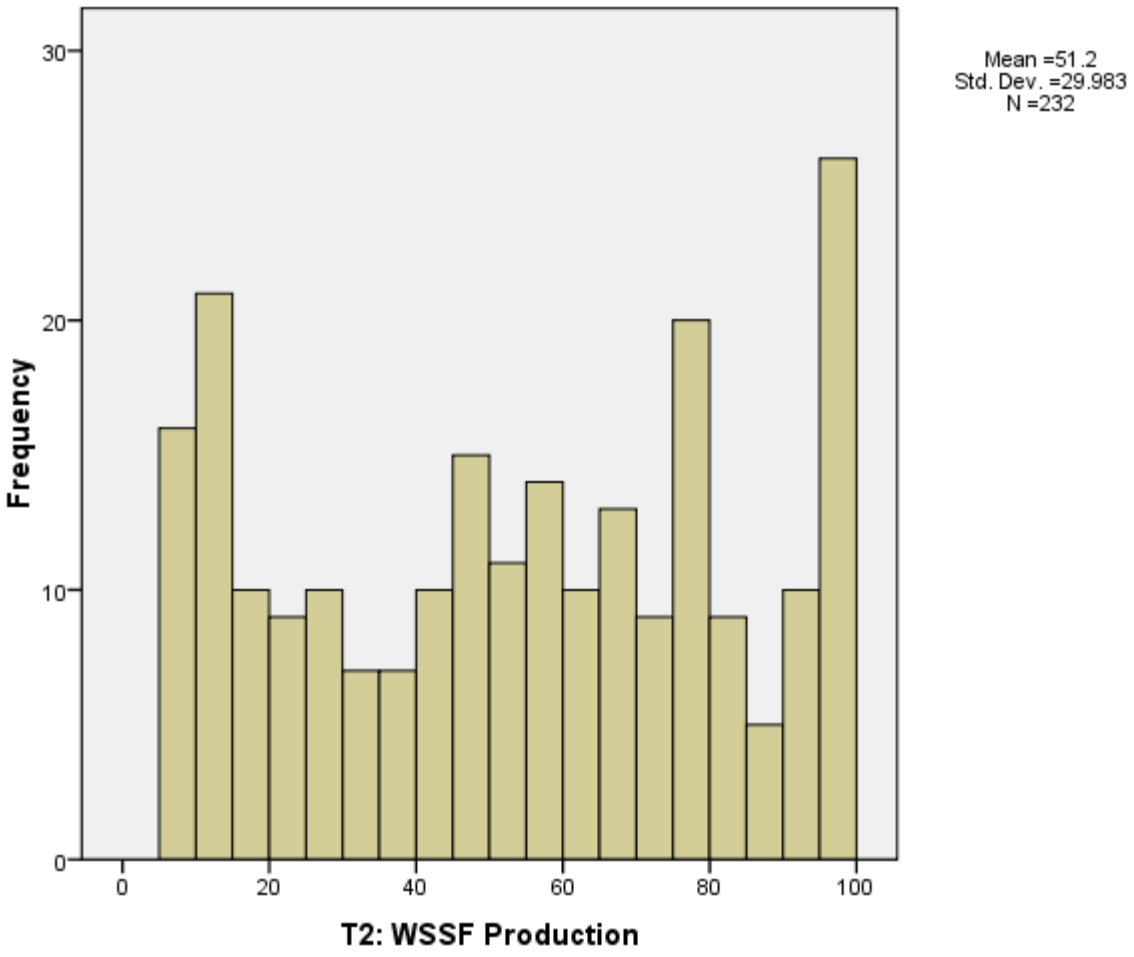

60
50
40
30
20
10
0
Sentence Complexity Scores
16
17
18
19
20
Age in Months
B10
M10
H10


Percentile Group
Bottom 10%
Middle 10%
Top 10%
Bottom 10%
Middle 10%
Top 10%
Fit line for Total
Scale
3.0
2.5
2.0
1.5
1.0
T2 WS Production %ile
T1 WG Production %ile
0
20
40
60
80
100


Percentile Group
Bottom 10%
Middle 10%
Top 10%
Bottom 10%
Middle 10%
Top 10%
Fit line for Total
Frequency Scale
7
6
5
4
3
2
1
T2: WS Production %ile
T1: WG Comprehension %ile
0
20
40
60
80
100


Percentile Group
Bottom 10%
Middle 10%
Top 10%
Bottom 10%
Middle 10%
Top 10%
Fit line for Total
Scale
4.0
3.5
3.0
2.5
2.0
1.5
1.0
T2: WS Production %ile
T1: WG Gestures %ile
0
20
40
60
80
100


Mean =51.21
Std. Dev. =33.018
N =212
Frequency
0
10
20
30
40
0
20
40
60
80
100
T1: WSLF Comprehension


Mean =51.2
Std. Dev. =29.983
N =232
Frequency
30
20
10
0
0
20
40
60
80
100
T2: WSSF Production