The criterion validity of four vocabulary tests compared with a language sample

Teresa A. Ukrainetz and **Carol Blomquist**
University of Wyoming

Abstract

This study provided empirical evidence of the validity of four vocabulary tests: Peabody Picture Vocabulary Test-III (PPVT-3), Expressive Vocabulary Test (EVT), Receptive One-Word Picture Vocabulary Test (ROWPVT), and Expressive One-Word Picture Vocabulary Test-Revised (EOWPVT-R) against a 150-utterance language sample for 28 normally-developing pre-school children. Systematic Analysis of Language Transcripts (SALT) standard scores on number of different words, total number of words, and mean length of utterance were used to provide evidence of convergent and discriminant criterion validity. Performance on the vocabulary tests showed significant weak to moderate correlations with the semantic measure, and the predicted lower relationship with the non-semantic measures. Despite this, individual score analysis showed considerable variation, indicating caution predicting conversational language performance from standardized tests.

Introduction

Vocabulary is commonly assessed with standardized, norm-referenced tests. It is important to know whether these measures portray children's vocabulary knowledge as it is manifested in daily life. This study examined the criterion validity of four vocabulary tests for the purpose of estimating the relative vocabulary knowledge of pre-school children. The empirical criteria used to evaluate validity were semantic and non-semantic measures calculated on a sample of language from activities common to a child's life.

Address for correspondence: Teresa Ukrainetz, Division of Communication Disorders, P.O. Box 3311, University of Wyoming, Laramie, WY 82071-3311, USA. E-mail: tukraine@uwyo.edu

 10.1191/0265659002ct227oa

Determining a vocabulary criterion

Validity concerns whether and how well a test measures what it is purported to measure (Anastasi, 1982). Degree of validity is dependent on the uses to which a test is put: a test may be more valid for a broad diagnostic distinction of 'plus or minus' language impairment than for a profile of language domain performance. Validity is a unitary construct, with the various sources of evidence converging toward a judgment on the extent and quality of validity (Messick, 1989). Validity is typically grouped into construct-, content-, and criterion-related, with evidence for each arising from logical and empirical sources (Hutchinson, 1996).

A logical analysis is generally the easier route through which to obtain evidence of validity. This approach involves examination of the test and comparison of the contents to the test-maker's or test-users' understandings of what ought to be in a test to measure a particular concept for a particular use. Informed test-users can often make many logically-based decisions about validity by consideration of the test structure, the items and foils, how test-takers respond to items, and how responses are scored. For example, a vocabulary test should reasonably contain words sampled from the lives of the group for which it is intended. There should be a range of word types, an absence of grammatical and contextual cues, and clearly interpretable pictures. Tests possessing such features could logically be considered to have some validity in measuring vocabulary performance. Stockman (2000) provides a recent example of such an analysis of the Peabody Picture Vocabulary Test-III (Dunn and Dunn, 1997).

Empirical validity is established through a data-based analysis of the test in relation to what it purports to measure. Performance on the test is compared to performance in other independently observable activities (Anastasi, 1982). While construct and content validity can be determined to some degree through logical analysis of the test and its administration, criterion validity inherently involves empirical evidence. The skills and concepts involved in standardized testing of 'vocabulary knowledge' are assumed to be stable, context-independent entities. However, they are also assumed to have some measurable manifestation in sufficient daily life situations, though not in as efficient or controlled a manner as through standardized testing (thus the reason for standardized testing). Criterion validity involves determining to what extent the standardized test results predict performance in these more difficult-to-measure life situations.

To determine the criterion validity of vocabulary tests for the purpose of estimating vocabulary knowledge, textbooks inform us that comparisons are made with an independent and acceptably valid measure of vocabulary.

However, as the foregoing suggests, the selection of a criterion is not a simple matter. Not surprisingly, the comparison most commonly made is with other norm-referenced standardized tests. The tests may be in the same domain or in related domains. This method provides an equally efficient and controlled source of data for comparison. However, data is not information and this easy comparison provides little information on the validity of these tests for describing vocabulary knowledge.

The first problem with this convenient test-to-test validation method is circularity. When tests are compared with other tests, one only has information that the test under examination is similar in some way to the other test. If the other test does not itself have acceptable evidence of validity, then little information on validity for the new test has been obtained (Anastasi, 1982).

The second problem is identifying the source of the similarity in performance. Do two tests show similar results because they are both measuring vocabulary or because they both measure test-taking performance or attentiveness or some other factor? It is even possible that two tests actually measure different constructs, such as vocabulary and intellectual performance, but show similar patterns of performance. In addition, sizeable raw score correlations may arise mainly due to developmental change: children generally get more items correct as they get older regardless of what is being measured.

What then would be more acceptable vocabulary criterion measures? Measures that might serve as criteria against which test performance could be compared include diary reports, vocabulary counts, and language sampling. These measures assess behaviour in a way closer to that of daily life performance. A record of all the words a child knows has the closest verisimilitude to daily life. However, for children beyond 3 years of age, the number of words known far surpasses the possibility of a comprehensive assessment. Another alternative is language sampling. Language sampling is like standardized testing in that it consists of a sample of what the child knows that is intended to represent the larger total knowledge possessed. It has the necessary element of being closer than standardized tests to daily life performance.

Language sampling as a criterion

In this study, language sampling was chosen as the criterion against which the vocabulary tests were compared. Language samples consist of the words a child uses in actual discourse situations. Language sampling is a cornerstone method of assessing children's communicative development in research and practice (Evans, 1996; Klee, 1985). The developmental course of the structure

and use of language has been charted based primarily on this approach. In addition, identification and evaluation of language disorders has involved examination of the child's performance in a conversational setting in addition to standardized testing.

One of the greatest strengths of language sampling is seen to be its direct representation of a person's language production system, as opposed to the secondary representation afforded by elicitation procedures such as standardized testing (Klee, 1985). Language sampling is sufficiently flexible to be fitted to individual children and their communicative situations. There are also generally accepted procedures for eliciting, transcribing and analyzing language samples, such as following the child's lead in conversation, typing one utterance per line, rules for calculation of mean length of utterance and determining syntactic complexity and age comparison information (e.g., Miller, 1981). In addition, current methods of language sample analysis allow calculation of standard scores for some measures (e.g., Miller and Nockerts, 2000).

Despite the developments in language sampling, it continues to be a complex method of assessment. Children's displays of vocabulary change with the discourse context. Interpretive judgements must sometimes be made in eliciting, transcribing, analyzing and interpreting language samples. In addition, language comprehension is difficult to determine in a naturalistic context. Finally, the selection of a pure vocabulary measure that has no influence from grammar or verbosity has been an ongoing challenge to researchers.

Test-to-test comparisons introduce circularity. Comprehensive vocabulary checklists are not possible beyond the toddler years. Language sampling introduces its own challenges, but the fairly direct representation of language in daily life and the empirical evidence available indicate it is a reasonable choice for contributing to our understanding of the criterion validity of vocabulary tests. The following section details some of the research supporting the use of language sampling and the particular measures that were used in this study.

Empirical validity of language sampling

Samples of language can be analyzed in many ways. Mean length of utterance (MLU), a measure of grammatical complexity, is a commonly used measure. This measure shows a strong correlation with age in normally developing children, children with specific language impairment, and children with learning difficulties (Klee, 1985; Miller, 1981; Templin, 1957). It also has substantial test–retest reliability (Gavin and Giles, 1996). MLU has been

shown to have clinical utility, separating children with language impairment from those with normally developing language (Aram *et al.*, 1993; Dunn *et al.*, 1996; Klee *et al.*, 1989; Restrepo, 1998).

An example of the diagnostic utility of MLU is demonstrated in Aram *et al.* (1993). Aram and colleagues evaluated discrepancy, deficit, and standardized operational criteria for identifying language impairment with 252 children aged 3–7 years with specific language impairment. The investigators reported considerable mismatch between clinical identifications and the criteria examined, which involved primarily standardized tests. The best single measure match between clinical identification and criteria was achieved using MLU transformed into standard scores based on norms from Miller (1981).

Dunn *et al.* (1996) further examined the utility of standardized test and language sample measures in 201 children clinically identified as having specific language impairment and 41 children considered as normal. The combination of MLU standard scores (using Miller and Chapman, 1985 data) with per cent structural errors and chronological age was the optimal subset of predicting identification of language impairment, even compared with the best psychometric discrepancy criteria.

There is less empirical information about other measures of language production. Klee (1992) analyzed the descriptive and diagnostic utility of nine language sample measures on 24 children with specific language impairment and 24 normally developing children aged between 2 and 4 years. The relevant measures for the current study were total number of words (TNW), number of different words (NDW) and type-token ratio (TTR).

TNW is an index of general language facility, reflecting 'a number of factors including speaking rate, length of utterance, speech motor maturation, utterance formulation ability, and word-retrieval efficiency' (Miller, 1981: pp 213–214). TNW also reflects volubility, which varies with personality and cultural factors (Crago, 1990), in addition to language facility. Klee (1992), along with other studies (e.g., Miller, 1981; Templin, 1957), found that TNW increases with age. Klee also reported that TNW showed diagnostic utility with significant separation between clinical groups on a regression analysis.

NDW, the number of different words obtained from language samples of a fixed number of utterances, is viewed as an index of semantic diversity (Miller, 1981; Klee, 1992). NDW is somewhat affected by utterance length, but it is considered to reflect primarily semantic factors. NDW has been shown to increase with age (Templin, 1957; Miller, 1981; Klee, 1992). Gavin and Giles (1996) reported high test–retest reliability for NDW on language samples of 150 utterances or more for preschool children. Klee (1992) found that NDW differentiated clinical groups to a similar extent as TNW.

Watkins *et al.* (1995) examined NDW for 25 pre-school children with specific language impairment, and their age and language matches. Watkins *et al.*, calculated NDW on both a set number of utterances and a set number of tokens with the latter controlling for differences in utterance length. NDW differentiated age-matched children from children with specific language impairment and younger age-matched children for both types of NDW measures.

Type-token ratio (TTR) has received attention as a measure of lexical diversity that accounts for the general volubility of the speaker. Its robustness, with lack of variation for age, sex, and social class, had been seen as a positive attribute for clinical purposes (Klee, 1992). However, several studies have raised concerns about the utility of TTR as a clinical measure. Klee reported that it failed to vary between children with language impairment and those who were normally developing. Watkins *et al.* (1995) reported a similar lack of differentiation between normally developing children and children with specific language impairment. Klee reported that modifications to TTR, involving a fixed number of words rather than utterances, still did not show consistent developmental change and failed to differentiate between normal and clinical groups. TTRs calculated on verbs only have shown better results, with lower verb TTRs for children with specific language impairment (SLI) than for age- and language-matched peers (Rice and Bode, 1993; Watkins *et al.*, 1993).

The current study

The purpose of the current study was to determine the degree to which two receptive and two expressive vocabulary tests show empirical evidence of validity when compared against the semantic measure NDW obtained on a 150-utterance language sample obtained across three contexts. Two other non-semantic language measures were calculated from the language sample, TNW and MLU, and compared with the tests to provide evidence of discriminant validity. In addition, modality, normative group, and task factors were investigated.

Method

Participants

Twenty-eight children (15 boys and 13 girls), from two local childcare centres in southeastern Wyoming, participated in the study. The children ranged in age

from 3;11 to 6;0 years, with a mean age of 4;9 years. Approximately 60 consent forms were sent out, to all the parents of 4- to 6-year-old children attending two area childcare centers. Thirty consent forms were returned. One participant was dropped due to unintelligibility in the language sample and one participant's data was lost due to technical difficulties, resulting in 28 final participants. Although no effort was made to limit the sample to normally developing children, there were no children with histories of language impairment, neurological, hearing, or emotional difficulties. The children came from families with a range of income levels from lower middle class to upper middle class. Most of the children were white, with one child of mixed black and white background.

Procedure

Four vocabulary tests were administered and a language sample was obtained over two 45-min individual sessions with each child by the second author, a graduate student in speech-language pathology. Each session consisted of two vocabulary tests and either a narrative sample or a conversation and expository sample. The vocabulary tests and the language samples were administered in a random order within and across sessions.

Standardized tests. The four vocabulary tests examined in the current study were the Peabody Picture Vocabulary Test-III (PPVT-3, Dunn and Dunn, 1997), the Expressive Vocabulary Test (EVT, Williams, 1997), the Receptive One-Word Picture Vocabulary Test (ROWPVT, Gardner, 1985), and the Expressive One-Word Picture Vocabulary Test-Revised (EOWPVT-R, Gardner, 1990). These tests were selected based on their common usage and the contrasts available: two receptive, two expressive, two developed as companion tests by one company and two by another. The tests were administered according to the manual instructions.

The PPVT-3 measured receptive single-word vocabulary. The examiner stated a word and the child pointed to one of four pictures that best represented the test word. The words were picturable nouns, verbs, or adjectives. The normative sample for this test was representative of the United States, stratified for geographic region, economic level, race and ethnicity, with approximately 100 participants at each age level.

The EVT measured expressive single-word vocabulary. The examiner pointed to a picture silently or with a word label. The child labelled the picture or provided a synonym for the word. The pictures were nouns, verbs, or adjectives. The EVT employed the same normative sample as the PPVT-3.

The ROWPVT measured receptive single-word vocabulary. Like the PPVT-3, it used a target picture and three foils. The normative group was drawn from the San Francisco Bay area. There was no report of demographic representation in the manual. There were approximately 30 participants in each of the 3–6-year age levels.

The EOWPVT-R measured expressive single-word test. The child labelled or provided a category name for the items represented. The normative group was drawn from the same San Francisco Bay area as the ROWPVT, but was not the same group. There were approximately 100 participants at each age level.

Language sample. The language sample consisted primarily of conversational discourse, with briefer narrative and expository discourse sections. The conversation context was play with a farm set. The farm set comprised six pairs of farm animals, a pumpkin patch, a hay bale, a tractor and trailer, and two farmers. The examiner used the same initial question (e.g., 'I wonder what the animals would do at the beginning of the day?') for each child to elicit conversational speech. The conversation then varied with the responses of the children, but an effort was made to elicit at least all the animal names.

The narrative section involved telling a story from a wordless picture book, *Frog Goes to Dinner* (Mayer, 1974). The examiner looked at the cover with the child and discussed what might happen in the story. She then asked the child to tell the story to the examiner. The prompts varied with the child's responsiveness, but the child's attention was drawn to the items depicted in each picture to provide similar vocabulary opportunities.

The expository portion involved description of two black and white line drawings from a therapy activity book (Heinze and Johnson, 1985). The picture presented a family in backyard activities and in house activities. An example of picture description was given by the examiner for a farmyard picture and then the child was asked to describe the two pictures. The child was prompted to describe items depicted in the picture if they did not occur spontaneously.

The language samples were transcribed, analyzed, and transformed into standard scores with the language analysis software Systematic Analysis of Language Transcripts (SALT, Miller and Nockerts, 2000). The language samples were initially approximately 170–250 child utterances in length across the three contexts. The samples were cut to a standard 150 utterances for each child: 100 from the conversation section during the farm play, 40 from the narrative section, and 10 from the expository section. The conversation

starting cut occurred after 10 child utterances. The narrative and expository utterances were taken from the beginning of the sections.

The samples were considered as conversation across three contexts rather than a combination of three distinct discourse genres. For these young children, the narrative and expository sections were conversational in tone with the examiner providing frequent prompts and affirmations. For the narrative, the children often described pictures rather than maintaining a story line. For the exposition, the children often labeled items in the picture rather than describing in a coherent text.

Three measures were calculated with SALT: Number of Different Words (NDW), Total Number of Words (TNW) and Mean Length of Utterance (MLU). NDW is a measure of semantic diversity and was arrived at by counting the number of different root word types in a sample. A fixed number of utterances was used rather than a fixed number of tokens despite the influence of utterance length because the utterance set allowed use of the SALT age-referenced standard scores. TNW is a measure of general verbal fluency that consisted of a count of all word tokens in a sample. MLU is a measure of grammatical complexity. It was arrived at on a t-unit basis, where utterances were terminated after each main clause and all attached subordinating clauses or non-clausal structures.

The SALT age-referenced database was used to obtain norm-referenced scores. For comparability with the test scores, a standard score with a mean of 100 and a standard deviation of 15 was calculated for the measures from the profiler results. The relevant portions of this database consisted of conversational language sample data from typically developing 3–6-year-old children from pre-schools and schools in the Madison metropolitan area and rural areas in northern Wisconsin. There were 15–35 children in each 12-month age group for a 150-utterance sample.

Criterion characteristics. Inter-rater agreement between the two transcribers was calculated on three children's language samples, chosen at random. Point-to-point agreement on word and utterance division matches were 87% and 88% respectively. All the test scorings were checked for accuracy by a second scorer.

A measure of internal consistency was calculated on two 60-utterance samples from the larger language samples for the 28 participants. Each sample comprised 35 conversation, 15 narrative, and five expository utterances, with the first sample involving the first utterance sets and the second sample involving the subsequent utterance sets. Pearson product–moment correlations were high–moderate: 0.79 for NDW, 0.71 for TNW, and 0.74 for MLU. This

Table 1 The per cent distribution of scores on the language sample measures compared to the normal curve distribution

	Standard deviation					
Measure	−3 to −2	−2 to −1	−1 to 0	0 to +1	+1 to +2	+2 to +3
Normal distribution	2	14	34	34	14	2
NDW	3	14	36	32	14	0
TNW	0	14	29	32	18	7
MLU	0	7	29	39	14	11

Note: Calculated based on an expected mean of 100 and a standard deviation of 15.

indicated reasonable consistency between early and later segments of the language sample.

The SALT standard score performance for this relatively small sample of children showed approximately normal characteristics for the main measure, NDW. NDW showed a mean standard score of 98.2, with a standard deviation of 13.6, close to the expected 100 and 15, respectively. The distribution of scores was approximately normal for NDW (Table 1). These results supported the appropriateness of use of this measure for this population of children.

TNW and MLU showed mean standard scores of 103.9 and 106.5 respectively. This was higher than the expected 100. Both distributions showed a negative skew with more scores than expected in the upper end of the distribution. Standard deviations were 13.6 and 14.7. The higher mean performance for MLU and TNW than NDW may have been due to the inclusion of narrative elements, which have longer utterances than conversation (Miller and Nockerts, 2000), in the conversational language sample.

Results

The purpose of this study was to provide empirical evidence of validity for four commonly used vocabulary tests against a language sample. Statview (SAS Institute, 1998) was used to perform all statistical analyses with the significance level set to $P < 0.05$ for all comparisons.

Correlational analysis

Pearson correlations were calculated for standard score performance across the tests and measures (Table 2). ROWPVT and NDW showed the strongest correlation ($r = 0.61$). The ROWPVT, EVT, and EOWPVT-R correlations

Table 2 Standard score correlations for tests and language sample measures

	NDW	TNW	MLU	PPVT-III	ROWPVT	EVT
TNW	.86					
MLU	.88	.99				
PPVT-III	.36	.12	.17			
ROWPVT	.61	.46	.51	.79		
EVT	.48	.25	.29	.75	.66	
EOWPVT-R	.46	.32	.36	.84	.79	.80

Note: $r > 0.36$ significant at $P < 0.05$; $r > 0.46$ significant at $P < 0.01$.

with NDW were significant at $P < 0.01$. The PPVT-3 showed the lowest correlation with NDW, at 0.36. This magnitude of correlation was significant at $P < 0.05$. The tests correlated with TNW and MLU at a lower level than with NDW. The ROWPVT again showed the strongest correlations ($r = 0.46$ and 0.51). The PPVT-3 correlations with TNW and MLU were the lowest, showing no reliable relationship ($r = 0.12$ and 0.17).

Consistent modality differences in standard score correlations were not demonstrated. The receptive test ROWPVT showed a higher correlation than did the two expressive tests with the expressive measure NDW. Similarity of normative group did not affect standard score performance relationship. The PPVT-3 was no more closely correlated to the EVT, constructed on the same normative group, than it was to EOWPVT-R, constructed on a different normative group.

Task factors had an effect on the pattern of correlations. Correlations were much stronger within a task than across tasks. NDW, TNW, and MLU were strongly correlated (0.86–0.99). Vocabulary test scores, across expressive and receptive modalities, were moderately to strongly correlated (0.66–0.84). The pattern of correlations for the PPVT-3 with the other three tests was no lower than the correlations among those three tests, despite the low correlations of the PPVT-3 with the language sample measures.

Distribution analysis

Mean standard score performance for the PPVT-3, ROWPVT, EVT, and EOWPVT-R ranged from 106 to 113 (Table 3). The tests showed higher means but similar standard deviations compared with NDW. The mean performance for the tests differed significantly from the mean NDW performance on a repeated measures analysis of variance, $F(4, 27) = 12.8$, $P < 0.0001$. Follow-up paired comparisons showed significant differences for all the tests compared with NDW. In addition, the

Table 3 Mean standard score performance for vocabulary tests

Test	Standard score mean	Standard deviation
PPVT-III	107.3	12.3
ROWPVT	106.1	13.9
EVT	108.8	12.2
EOWPVT-R	113.2	15.9

Table 4 The per cent distribution of scores on the vocabulary tests compared with the normal curve distribution

	Standard deviation					
Measure	−3 to −2	−2 to −1	−1 to 0	0 to +1	+1 to +2	+2 to +3
Normal distribution	2	14	34	34	14	2
PPVT-III	0	3	11	50	29	0
ROWPVT	0	7	21	50	18	3
EVT	0	0	21	46	25	7
EOWPVT-R	0	7	18	21	43	7

Note: Calculated based on an expected mean of 100 and a standard deviation of 15.

EOWPVT-R was significantly higher even than the next closest test mean, EVT, $t(27) = 2.4$, $P = 0.02$.

The tests showed an approximately bell-shaped curve, but displaced upward (Table 4). The PPVT-3, ROWPVT, and EVT showed 46–50% of the sample scoring in the first standard deviation above a mean of 100, instead of the expected 34%, and 25–29% in the second standard deviation, instead of 14%. The EOWPVT-R showed a large negative skew, with the mode of 43% occurring in the second standard deviation above the mean.

Individual performance analysis

To estimate the practical significance of the relationships identified in the correlational analysis, individual performance was examined. A comparison of absolute scores showed considerable differences between NDW and the tests. All four of the children who did poorly on NDW (standard scores of 65–83) scored 20–27 points higher on the ROWPVT, with scores from 92 to 106. They scored 21–43 points higher on the EVT, 25–39 points higher on the EOWPVT-R, and 28–37 points higher on the PPVT-3. The four highest NDW scores (119–122) showed no discernable relationship with the PPVT-3, EVT, and EOWPVT-R test scores: the points ranged from seven points lower to 14 points higher. There was reasonably good agreement between NDW and ROWPVT, with scores differing by only two points lower to three points higher.

To circumvent the differences in mean scores, individual rankings were compared for the top four and bottom four scorers on NDW. The lowest scorer on NDW was identified as one of the four lowest scorers on the ROWPVT, EVT, and EOWPVT-R. None of the other three low NDW scorers were among the four lowest on the four tests. Each of the four tests identified only one of the four highest NDW scorers, with only the EOWPVT and the PPVT-3 identifying the same scorer.

Discussion

Relationships between language sample and tests

Results showed significant weak-to-moderate positive correlations between each of the tests and the semantic measure NDW. This indicated that the four vocabulary tests are measuring some aspect of semantic knowledge and thus have evidence of criterion validity.

Further support for criterion validity came from the comparison of the tests to the non-semantic language measures TNW and MLU. A lower correlation was expected for TNW and MLU than for NDW. This pattern was obtained, providing evidence that the vocabulary tests were measuring something more like vocabulary and less like verbal fluency or grammar. All these language measures were expected to show some positive correlation because they all appear to sample language. The use of standard scores rather than raw scores provided some partialling out of age, so the similarities were based more on vocabulary variation rather than simply developmental change. This supported the usefulness of the tests in describing relative vocabulary knowledge within an age group.

The PPVT-3 consistently produced the lowest correlations with the language sample measures. While the NDW to PPVT-3 correlation was significant, it was low, and the TNW and MLU relationships were almost non-existent. There was no discernable reason for this pattern of results. This test appears to sample vocabulary knowledge, is carefully constructed, and has a long history of use with prior versions (Stockman, 2000; Ukrainetz and Duncan, 2000). Reports have been made about the higher standard scores obtained on the PPVT-3 compared with the previous edition by several investigators (Stockman, 2000; Ukrainetz and Duncan, 2000; Washington and Craig, 1999), but no other issues specific to the PPVT-3 have been reported. The small sample size of the current study prevents firm conclusions, but these results suggest some concern with criterion validity and warrant further investigation.

Modality differences did not affect the pattern of correlations. Expressive tests were no more similar to the expressive language sample than were the

receptive tests. In addition, cross-modality vocabulary tests were as strongly correlated as within-modality tests. This is not a reflection on absolute numbers of items known – receptive vocabulary is certainly larger than expressive vocabulary – but on relative performance. The children in this sample who had high standard test scores expressively also tended to have high standard test scores receptively. These results do not fully answer whether both modalities need be assessed clinically, because the sample was composed of children considered to be developing normally. Children with expressive language impairments would be expected to show larger differences between vocabulary modalities.

The pattern of correlations did not show differences based on normative groups. Tests constructed on the same normative group – the PPVT-3 and the EVT – were no more similar than tests constructed on different normative groups – the PPVT-3 and the EOWPVT-4. Similarity in pattern of response occurred despite mean score differences between the EOWPVT-R and the other tests. This is positive because low scorers on one test will likely be low scorers on another test, even if the absolute scores vary considerably. Clinicians can deal with this by having differing expectations: low scores on the EOWPVT-R might be below 92 and low scores on the ROWPVT might be below 85 (although precise score differences would need to be confirmed on a larger sample).

Distribution comparisons

Mean score differences were demonstrated between NDW and the tests. Compared with NDW, the tests showed significantly higher mean standard scores. This difference may relate to demographic patterns—vocabulary is experientially sensitive, and Wyoming experiences may be more similar to Wisconsin experiences than those of San Francisco (ROWPVT and EOWPVT-R) or groups sampled across the country (PPVT-3 and EVT).

A more probable reason for the higher test standard scores is the difference in the ability make-up of the normative groups. McFadden (1996) outlines a concern with norm-referenced testing when normative groups are composed of only normally developing children. Such tests eliminate low scorers from the normative group, resulting in a distribution with the lower end missing. Such practices lead to lower standard scores for children taking the test. The mean standard score performance in this study fits such an explanation.

The vocabulary tests use a full range of abilities, including low scorers, in their normative groups. The normal-only sample of the current study would be expected to perform slightly higher than the average performance of the full

range normative group. This occurred. In contrast, the SALT normative group consists of only normally developing children. When the study sample was compared with SALT, the mean performance was at the expected level of 100 compared with the SALT normal-only normative group. The difference in normative groups reflects negatively on SALT because of the undesirable consequences, such as identifying normal children as language impaired, resulting from normal-only normative groups (McFadden, 1996).

The EOWPVT-R was a concern because the mean performance on it was higher than the other tests as well as than NDW. There was no apparent reason for the much higher mean standard score for the EOWPVT-R, other than the normative group for this test being exceptionally low performing. The higher mean standard score performance on the EOWPVT-R will affect language profiling even if it is restricted to comparison with other tests. A discrepancy analysis can only be carried out if the normative groups perform similarly (Anastasi, 1982). When performance on the expressive EOWPVT-R was compared with its companion receptive test, ROWPVT, a gap of more than one standard deviation (e.g., 50th percentile versus 16th percentile) in favour of expressive skills occurred for five of the 27 children. An EVT and PPVT-3 comparison produced only one child with a gap of more than one standard deviation, resulting in a large difference depending on which test pair is used. The two expressive tests differed by more than one standard deviation for seven of 28 children, making conclusions very different depending on which test is used. Further investigation of this difference is warranted.

Individual result analyses

Analysis of individual high and low scorers showed, despite the significant and sizeable correlations, that there was little agreement between NDW and the tests. The differences in mean scores made a direct comparison between NDW and the tests problematic. Three of the tests were approximately nine points higher than NDW. A nine-point difference would represent agreement on an individual basis between NDW and three of the tests. However, differences were much larger than that for the low scorers, particularly on the PPVT-3, EVT, and EOWPVT-R. The situation was less predictable for the high scorers on those three tests, with scores both lower and higher than NDW, making any kind of systematic compensation impossible. The ROWPVT showed the best agreement, with scores 20–27 points higher for the low scorers and almost identical for the high scorers. A comparison of the four highest and four lowest scorers regardless of absolute score supported this lack of predictable

relations. Seventy-five per cent of low or high scorers on the NDW would not be identified as such with any of the vocabulary tests.

There is clearly a relationship between test and NDW performance. However, this degree of variation is an issue. Several studies have already indicated concerns about the diagnostic validity of language tests in general (Aram *et al.*, 1993; Dunn *et al.*, 1996), and vocabulary tests specifically (Gray *et al.*, 1999). In making decisions, speech-language pathologists typically use a combination of standardized and descriptive measures. This converging evidence approach is a good insurance against gross errors in overall description and diagnosis, although specific domain descriptions, where there is less testing redundancy, are still a concern.

Activity and skill: reflections on what is being measured

The daily life activities used for language sampling were very different from the activity of 'test-taking.' Not surprisingly, there were stronger correlations between language sample measures and between test measures than across the two activities. The correlations present between the two activities were an indicator of the 'vocabulary knowledge' that was being measured independent of context. This is consistent with conventional structural approaches to language development. However, more recent functional theories of language development may be useful in guiding evaluation of our language assessment measures (Evans, 1996; Ukrainetz, 1998).

Traditional assessment practices have tended to break language performance into its component skill domains. Over the decades, the domains have multiplied, from 'syntax and vocabulary' to areas such as lexical semantics, relational semantics, meta-semantics, narrative structure, conversational pragmatics, and expository comprehension, to name a few. Evans (1996) suggests this proliferation has been due, in part, to the variability of results across contexts and the resultant need to examine many aspects of language in many contexts.

Evans (1996) suggests that language performance variability should not be considered an undesirable by-product, but rather an intrinsic and positive part of language performance, reflecting the child's adaptability to long-term and momentary situational variation. As such, skills and understandings are intrinsically related to the context in which they are practised. Differences are not 'errors' or 'noise' but are part of the knowledge itself. The way one manifests one's understanding of *chair* in a conversational setting is very different from the manifestation in a structured picture labelling situation. Neither manifestation is more 'real' or more 'correct.' Evans proposes shifting from a structural to this functional understanding of

language performance, with the intriguing but yet speculative recommendation of dynamical systems.

Another functional approach that is more immediately applicable is suggested by Ukrainetz (1998). Ukrainetz suggests applying Soviet activity theory and its Western developments (Rogoff, 1990; Rogoff and Lave, 1984; Wertsch, 1981) to language intervention. This social-interactionist theory elevates 'activity' over 'skill' as the basis of development. It suggests that mental activity not only grows out of, but reflects the structure of everyday, practical activity that involves both the individual and his or her environment (Wertsch, 1981).

In an activity approach to assessment, language performance is broken down into major contexts of use or significant daily life activities such as peer conversation, standardized test-taking, or writing history compositions. Judgments of language competence are restricted to the situations assessed and other functionally similar situations. As such, good test-taking vocabulary might be expected to be reflected in worksheet assignments, but not in social communicative exchanges or a written imaginative narrative. Validation of a vocabulary test would involve comparing performance on the test with other daily life activities that are like the test. Differences in performance between very different activities would be expected.

This approach to assessment would be more consistent with current naturalistic intervention approaches involving teaching language where, how, and for the reasons it actually happens. However, it does involve a reconceptualization of standardized testing as a context in itself—these tests were originally designed to sample 'pure' isolated language skills independent of context. In this alternate functional view, test-taking itself is an activity. Its closest counterparts would be the other standardized testing that occurs in school and the worksheets and test-training that occur in preparation for these tests. These are significant daily life activities in themselves, so performance on a standardized test would provide useful information on how children could be expected to perform on classroom testing-related activities. Standardized testing would not be expected to predict performance in daily life activities that are more complex, interactional, or self-directed such as negotiating with mother for a sleepover or participating in a class project on the life-cycle of butterflies.

Conclusions

This small sample study provided some empirical evidence that the EVT, the ROWPVT, and the EOWPVT-R have a reasonable degree of criterion validity.

These three tests showed significant positive moderate correlations with SALT standard scores on the semantic measure, NDW, from a conversational language sample for 28 children aged 4–6 years. All of the tests showed discriminant validity by weaker positive correlations with the more indirectly related measures of total number of words and mean length of utterance. The PPVT-3 showed some validity with significant weak positive correlations with NDW and lower correlations with the non-semantic language measures. However, the low level of correlations of the PPVT-3 with the language sample measures raised concern about this test.

All the tests showed higher mean standard score performance than NDW, possibly due to differences in normative group ability composition. The very high mean performance on the EOWPVT raised particular concerns about that test. Analysis of individual scores revealed a high rate of mismatches between test and language sample performance despite reasonable correlations and after taking into consideration mean score differences.

These findings show that vocabulary tests are related to daily life vocabulary performance, but that there are significant caveats in using them. More fundamentally, it is suggested that assessment practices be shifted from the structural approach of examining underlying skills to a more functional approach, wherein competence is considered inextricable from the life activities in which it occurs. With this activity approach, children would be assessed on measures that are similar to their actual contexts of performance. Measures would be validated against criteria that provide similar contexts of use. This approach would alleviate the 'which is right?' puzzle when considering mismatches between results from different assessment methods such as standardized tests and language sampling. It would also better inform clinicians about expected language performance.

References

Anastasi, A. 1982: *Psychological testing*. NY: Macmillan.

Aram, D. M., Morris, R. and Hall, N. E. 1993: Clinical and research congruence in identifying children with specific language impairment. *Journal of Speech and Hearing Research* **36**, 580–91.

Crago, M. 1990: Development of communicative competence in Inuit children: implications for speech-language pathology. *Journal of Childhood Communication Disorders* **13**, 73–83.

Dunn, L. M. and Dunn, L. M. 1997: *Peabody Picture Vocabulary Test-III*. Circle Pines, MN: American Guidance Service.

Dunn, M., Flax, J., Sliwinski, M. and Aram, D. 1996: The use of spontaneous language measures as criteria for identifying children with specific language impairment: an attempt to reconcile clinical and research incongruence. *Journal of Speech and Hearing Research* **39**, 643–55.

Evans, J. L. 1996: Plotting the complexities of language sample analysis. In: Cole, K. N., Dale, P. S. and Thal, D. J., editors, *Assessment of communication and language*. Baltimore, MD: Brookes, 207–56.

Gardner, M. F. 1985: *Receptive One-Word Picture Vocabulary Test*. Novato, CA: Academic Therapy.

Gardner, M. F. 1990: *Expressive One-Word Picture Vocabulary Test – Revised*. Novato, CA: Academic Therapy.

Gavin, W. J. and Giles, L. 1996: Sample size effects on temporal reliability of language sample measures of preschool children. *Journal of Speech and Hearing Research* **39**, 1258–62.

Gray, S., Plante, E., Vance, R. and Henrichsen, M. 1999: The diagnostic accuracy of four vocabulary tests administered to preschool-aged children. *Language, Speech, and Hearing Services in Schools* **30**, 196–206.

Heinze, B. A. and Johnson, K. L. 1985: *Easy does it: fluency activities for young children*. Moline, IL: Linguisystems.

Hutchinson, T. A. 1996: What to look for in the technical manual: twenty questions for users. *Language, Speech, and Hearing Services in Schools* **27**, 109–21.

Klee, T. 1985: Clinical language sampling: analysing the analyses. *Child Language Teaching and Therapy* **1**, 182–7.

Klee, T. 1992: Developmental and diagnostic characteristics of quantitative measures of children's language production. *Topics in Language Disorders* **12**, 28–41.

Klee, T., Schaffer, M., May, S., Membrino, I. and Mougey, K. 1989: A comparison on the age-MLU relation in normal and specifically language-impaired preschool children. *Journal of Speech and Hearing Disorders* **54**, 226–33.

Mayer, M. 1974: *Frog goes to dinner*. NY: Dial Books.

McFadden, T. U. 1996: Creating language impairments in typically-achieving children: the pitfalls of 'normal' normative sampling. *Language, Speech, and Hearing Services in the Schools* **27**, 3–9.

Messick, S. L. 1989: Validity. In Lin, R. L., editor, *Educational measurement, third edition*. NY: Macmillan.

Miller, J. 1981: *Assessing language production in children: experimental procedures*. Baltimore, MD: University Park.

Miller, J. and Chapman, R. 1985: *Systemic analysis of language transcripts*. Madison, WI: Language Analysis Laboratory.

Miller, J. and Chapman, R. 1981: The relation between age and mean length of utterance in morphemes. *Journal of Speech and Hearing Research* **24**, 154–61.

Miller, J. and Nockerts, A. 2000: *Systematic analysis of language transcripts*, v6.1. Madison, WI: Language Analysis Laboratory.

Restrepo, M. A. 1998: Identifiers of predominantly Spanish-speaking children with language impairment. *Journal of Speech, Language, and Hearing Research* **41**, 1398–411.

Rice, M. L. and Bode, J. V. 1993: GAPS in the verb lexicons of children with specific language impairment. *First Language* **13**, 113–31.

Rogoff, B. 1990: *Apprentice in thinking: cognitive development in social contexts*. New York, NY: Oxford University Press.

Rogoff, B. and Lave, J. 1984: *Everyday cognition: its development in social contexts*. Cambridge, MA: Harvard University Press.

SAS Institute, 1998: *Statview*. Cary, NC: SAS Institute.

Stockman, I. J. 2000: The new Peabody Picture Vocabulary Test-III: an illusion of unbiased assessment? *Language, Speech, and Hearing Services in Schools* **31**, 340–53.

Templin, M. C. 1957: *Certain language skills in children*. Minneapolis, MN: University of Minnesota Press.

Ukrainetz, T. A. 1998: Beyond Vygotsky: what Soviet activity theory offers naturalistic language intervention. *Journal of Speech-Language Pathology and Audiology* **22**, 122–33.

Ukrainetz, T. A. and Duncan, D. S. 2000: From old to new: concerns with score increases on the PPVT-III. *Language, Speech, and Hearing Services in Schools* **31**, 336–9.

Washington, J. and Craig, H. 1999: Performances of at-risk, African-American preschoolers on the Peabody Picture Vocabulary Test-III. *Language, Speech, and Hearing Services in the Schools* **30**, 75–82.

Watkins, R. V., Kelly, D. J., Harbers, H. M. and Hollis, W. 1995: Measuring children's lexical diversity: differentiating typical and impaired language learners. *Journal of Speech and Hearing Research* **38**, 1349–55.

Watkins, R. V., Rice, M. L. and Moltz, C. C. 1993: Verb use by language-impaired and normally developing children. *First Language* **13**, 133–43.

Wertsch, J. V. 1981: *The concept of activity in Soviet psychology*. NY: M. E. Sharpe.

Williams, K. T. 1997: *Expressive Vocabulary Test*. Circle Pines, MN: American Guidance Service.