Assessing Receptive Vocabulary in Small-town Canadian Kindergarten Children: Findings for the PPVT-R

Évaluation du nombre de mots compris chez les enfants d'école maternelle de petites villes canadiennes : conclusions relatives au PPVT-R

by • par

Peter Flipsen Jr., MA, S-LP(C)

University of Wisconsin-Madison
Madison, Wisconsin

ABSTRACT

The current study was a follow-up to a study by Flipsen (1993) that examined the use of the ROWPVT with small-town Canadian kindergarten children. In the current study, the PPVT-R was administered in the same community to 74 kindergarten children as part of a district-wide speech and language screening program. Results indicated a normal distribution of standard score performance using three different procedures. In addition, the number of screening failures (scores below a standard score of 80) met expectations from a normal distribution. Finally 78% of the children achieved basal scores at or above expected levels. Overall these findings suggest that the PPVT-R is an appropriate instrument for assessing receptive vocabulary in small-town Canadian kindergarten children.

ABRÉGÉ

La présente étude constituait un suivi de celle de Flipsen (1993) qui examinait l'usage du *ROWPVT* chez les enfants d'école maternelle de petites villes canadiennes. Dans la présente étude, le *PPVT-R* a été administré dans la même collectivité à 74 élèves de maternelle dans le cadre d'un programme de dépistage orthophonique de district. Les résultats ont démontré une répartition normale de scores standard, à partir de trois méthodes distinctes. En outre, le nombre d'échecs de dépistage (scores inférieurs au standard de 80) a reflété les résultats attendus d'une répartition normale. Soixante dix-huit p. 100 des enfants ont obtenu des scores de base égaux ou supérieurs aux cotes escomptées. Dans l'ensemble, ces résultats indiquent que le PPVT-R est un instrument approprié pour l'évaluation du nombre de mots compris chez les enfants d'école maternelle de petites villes canadiennes.

KEY WORDS

PPVT-R • standardized testing • validity • kindergarten • receptive language • Canadian children • vocabulary

The current study is a follow-up to a report by Flipsen (1993) which indicated significant concern with using the Receptive One-Word Picture Vocabulary Test (ROWPVT; Gardner, 1985) with small-town Canadian kindergarten children. In the current study, similar procedures were utilized to evaluate the application of the Peabody Picture Vocabulary Test - Revised (PPVT-R, Form M; Dunn & Dunn, 1981) with the same population.

Flipsen (1993) found that, despite a relatively large sample size (176), group performance on the ROWPVT was not normally distributed. The lack of normality was confirmed by the fact that the test also yielded a smaller proportion of screening failures than might be expected. As well, the study indicated that basal scores were being established at much lower levels than predicted necessitating longer and thus less efficient administration. The 1993 study was conducted as part of a kindergarten speech and language screening project, in a small resource-based town (population ~ 4000) in northern British Columbia during the 1991-92 and 1992-93 school years. Just prior to the 1993-94 school year, in response to the above concerns with the ROWPVT, the receptive vocabulary measure for the screening program was changed to the PPVT-R.

The primary purpose of the current study was to determine whether, unlike the ROWPVT, the performance of small-town Canadian kindergarten children on the PPVT-R was distributed similarly to the normative population. Such a similarity in distribution between the current sample and the PPVT-R's normative population would suggest that the two populations are similar enough in overall linguistic experience for individual test scores to be indicative of relative level of functioning in the domain sampled by the PPVT-R (i.e., receptive vocabulary).

The PPVT-R was selected as an alternative to the ROWPVT for the screening program for two reasons. First, the test continues to be widely used by many clinicians and in many research applications (e.g., Baker, Kummer, Schultz, Ho, & Gonzalez del Rey, 1996; Dawson, Blamey, Dettman, Barker, & Clark, 1995; Eskenazi & Trupin, 1995; O'Callaghan, Williams, Anderson, Bor, & Najman, 1995; Sturner, Layton, Evans, Heller, Funk, & Machon, 1994; Washington & Craig, 1992). Second, the PPVT-R was standardized on a sample of 4200 individuals representing a broad geographic cross-section of the United States. The standardization sample was a particular concern given that the problems with the application of the ROWPVT were thought to have been, in part, the result of its standardization sample (n = 1128) having been drawn exclusively from the San Francisco area. The larger sample and broader geographic base of the PPVT-R norms suggested a greater likelihood that it would be applicable to small-town Canadian kindergarten children. An empirical examination of the applicability of the PPVT-R to the current population appeared warranted however, in light of the findings of Flipsen (1993).

Method

Children Tested

The PPVT-R was administered to all children entering kindergarten in the community during the 1993-94 school year and thus the sample represented the entire cohort of children of kindergarten age during that year. Unlike the ROWPVT study, the current study group represented only a single year of screening (rather than two years with the ROWPVT), and thus involved a smaller sample size of 74. Age and gender data are shown in Table 1.

The children were largely of white lower-middle class and middle class background with approximately 5% being of Native descent. While no formal measure of socioeconomic status (SES) was made on any of the children or their families, SES status of the group was estimated from the makeup of the community at large. This procedure was felt to be reasonable given that the sample represented the complete cohort. The most relevant census data (Statistics Canada, 1991) indicated that, of 1275 households (total population 3804), 79.6% had annual incomes below $70,000 (29.8% below $30,000). As well, 61% of the 2510 adults in the community had no post-secondary education and fewer than 10% had attended university (with fewer than 4% completing university degrees).

Procedures

With the exception of the change from the ROWPVT to the PPVT-R, all procedures in the screening program were identical to those reported in Flipsen (1993). All administrations were conducted by the author who was serving as the school district's speech-language pathologist at the time. Testing took place within each child's kindergarten classroom during ongoing classroom activity. Following test administration, raw scores were converted to standard scores using the normative tables provided in the test manual.

The evaluation of the performance of the PPVT-R was accomplished in three ways. First overall performance trends were examined to determine whether the standard score distribution was similar to that of the normative population described in the test manual. A sample of this size would be expected to yield a normal distribution of scores. This evaluation was accomplished using three procedures. First overall mean and standard deviation of the scores was examined. Second, the nonparametric X^2 test of proportions across standard deviation categories was conducted as in the previous report. And third, the parametric Anderson-Darling test of normality was applied to the standard scores. This and all other statistical analyses were carried out using the software package MINITAB (Release 10Xtra; Minitab Inc, 1995).

The second test of the applicability of the PPVT-R to the current population involved examining the number of screening failures obtained. A screening failure was defined here as any standard score below 80, a criterion which also corresponded to scores falling below approximately the 10th percentile. As such it was expected to capture approximately 9-10% of the population. While this second test was not totally orthogonal to the test of the normality of the distribution (i.e., a normal distribution would very likely yield the corresponding number of screening failures), this test served as a more direct look at the sensitivity of the instrument with the study population for screening purposes.

Standard scores were chosen (as opposed to raw scores) for the current analysis primarily because of their ability to compensate for normal development (i.e., age differences) in the current study group (Bailey & Wolery, 1989, p. 30). Despite the narrow age range of the children in the current study (61-72 months), the test itself suggests development across the range. For example, there are two different designated starting points for testing children of kindergarten age. As well, identical raw scores translate to different relative levels of performance depending on the specific age of the child. For example, two children in the current study achieved a raw score of 64. For one child, age 61 months, this raw score translated to a standard score of 107 (almost one-half of a standard deviation unit above the mean), while for the second child, age 71 months, it translated to a standard score of 95 (one-third of a standard deviation unit below the mean). While both of these standard scores were within the normal range, the difference between them (8/10 of a standard deviation unit) illustrates the need to control for age differences. The ability of standard scores to compensate for development is further illustrated by a statistically significant Pearson correlation coefficient between raw score and age for the current study group (r = 0.29, F = 6.30, df = 73, p = 0.014) in contrast to a non-significant Pearson correlation between standard score and age (r = 0.023, F =0.04, df = 73, p = 0.846).

The third test of the PPVT-R involved examination of the efficiency of test administration, again as in the previous report, by calculating the median item number at which basal scores were established and then examining the number of children achieving basal at lower than expected levels. A lower basal score than expected would have indicated the need to reverse direction during testing increasing administration time for each individual and thus reducing efficiency of the screening program in general.

Results

Standard Score Distribution

Results for standard scores are shown in Table 1. Based on the mean and standard deviation data, the standard score distribution appeared to be very similar to that reported in the test manual (M = 103.4; *SD* = 14.8 vs. M = 100; *SD* = 15).

Table 1. Means, Standard Deviations (*SD*), and Ranges for Children's Ages and Obtained PPVT-R Standard Scores.

	Age (in months)				Standard Score		
	n	Mean	*SD*	Range	Mean	*SD*	Range
Males	38	67.1	3.6	61-72	103.2	14.9	54-137
Females	36	66.6	3.4	61-72	103.6	14.8	70-135
Total	74	66.9	3.5	61-72	103.4	14.8	54-137

A similar result was obtained when the distribution of standard scores (see Table 2) was examined. Applying the X^2 test of proportions, with the expected distribution using standardscore categories derived from areas under the normal curve (Triola, 1980), an X^2 value of 2.371 (5 df, p = 0.796) was obtained. This finding again indicated that the distribution of standard scores was not significantly different from normal. Finally, applying the Anderson-Darling test of normality resulted in an A^2 value of 0.454 (p = 0.264). This outcome also suggested that the distribution of scores was not significantly different from that expected with a normal distribution.

Table 2. Obtained Versus Expected Distribution of Scores Expressed as Number of Children.

Range	Males	Females	Overall	Expected
Below 70	1	0	1	2
70-84	2	4	6	10
85-100	14	8	22	25
100-115	12	18	30	25
116-130	8	5	13	10
131 +	1	1	2	2
Total	38	36	74	74

Note: Expected scores based on areas under the normal curve (Triola, 1980).

Screening Failures

As noted above, a screening failure was deemed to be a standard score below 80 (i.e., below the 10th percentile). Given a normal distribution, one would expect about 9-10% (7-8 children) to score below 80. Results indicated that 4/74 (5.4%) of the children scored below this criterion. To determine if this difference was statistically significant, another X^2 test of proportions was carried out. Results indicated no significant difference (X^2 = 0.884, 1 df, p = 0.347) between obtained and expected values.

Basal Score

The recommended starting items given in the test manual and the predicted basal items are shown in Table 3. Assuming that the recommended starting point for testing is a reasonable one, it was predicted that a normally developing child would establish a basal at least eight items

Table 3. Predicted Basal Item.

Age Range	Children Tested			Starting Item [1]	Predicted Basal Item [2]
	M	F	%		
2;6-3;5				1	9 +
3;6-3;11				10	18 +
4;0-4;5				15	23 +
4;6-4;11				20	28 +
5;0-5;5	15	13	37.8	30	38 +
5;6-6;0	23	23	62.2	35	43 +

Notes: 1. as per test manual. 2. highest of 8 consecutive correct.

above this point. A basal is established on the PPVT-R by identifying the highest set of 8 consecutive correct responses. Failure on any item in the first 8 requires the examiner to reverse direction in testing until the child identifies 8 consecutive items correctly. The examiner then returns to the forward direction to establish a ceiling.

The PPVT-R includes two possible starting points for children of kindergarten age (depending on their specific chronological age). The number of children meeting criterion for each of these two starting points is also shown in Table 3 (along with gender breakdown). As indicated, 37.8% (28) of the children were in the 5;0-5;5 age range and started at item 30 and 62.2% (46) were in the 5;6-6;0 range and started at item 35.

The median item number at which basal was established was calculated for the 74 children and results are shown in Table 4. Also shown is the number of children who achieved basal scores below the various item level categories.

Table 4. Obtained Basal Data Expressed in Terms of Cumulative Number of Children.

	5;0-5;5		5;6-6;0		Total	
Median	42		57		50.5	
	M	F	M	F	M	F
# above Item 43	7	7	19	19	26	26
# below Item 43	8	6	4	4	12	10
# below Item 38	4	4	2	2	6	6
# below Item 28	3	1	0	2	3	3
# below Item 23	2	0	0	1	2	1
# below Item 18	1	0	0	0	1	0

Note: Median equals item at which basal score established (highest of 8 consecutive correct).

Assuming that the younger children (aged 5;0-5;5) should have achieved basal at item 38 or above and the older children (aged 5;5-6;0) should have achieved basal at item 43 or above, the median basal data in Table 4 indicates that more than half of the children met this criterion. In fact, the specific data in Table 4 shows that only 22% (16/74; 8 per age category) of the children achieved basal at a lower level than expected.

Discussion

With a relatively large sample of children with a normal range of abilities, one would expect a normal distribution of performance. Using the type of standard scores employed with the PPVT-R, such a distribution would be expected to yield a mean score near 100 with a standard deviation near 15. The values obtained herein were very similar (103.4 & 14.8) to those expectations, suggesting that the scores obtained in the current sample were normally distributed. As was shown by Flipsen (1993) however, applying this criterion alone is insufficient for answering the question of the normality of the distribution. Using both the nonparametric X^2 test of proportions and the parametric Anderson-Darling test of normality confirmed a normal distribution of scores. This result suggests that, unlike the ROWPVT, the population of small-town Canadian kindergarten children in the current study is very similar in its distribution of performance to that of the normative population of the PPVT-R. A sample large enough to provide a normal distribution of abilities resulted in a normal distribution of performance and suggested that individual scores derived for this population using the PPVT-R would likely provide an adequate sample of the language domain tested by the PPVT-R. The question of the content validity of the PPVT-R is beyond the scope of the current investigation.

The failure to find significant differences between the current sample distribution and that of the normative population does raise two important sampling questions. The first is whether it is appropriate to use standard scores rather than raw scores. As was discussed previously, standard scores were used here because of the need to compensate for age. Since it might be argued that conversion to standard scores might have changed the essential character of the distribution (i.e., from non-normal to normal), the Anderson-Darling test of normality was applied post-hoc to the raw scores for the current study group and the same non-significant result ($A^2 = 0.584$, $p = 0.124$) was obtained. This same test was then applied to the raw scores from Flipsen (1993), confirming the non-normality of that distribution ($A^2 = 1.192$, $p = 0.004$). Thus in both studies, identical findings were obtained regardless of whether raw scores or standard scores were used.

The ability to obtain a non-normal distribution using standard scores is also supported by the findings of another study specifically involving the PPVT-R. Washington and Craig (1992) tested 105 low-income, African American children in kindergarten and first grade, all of whom spoke Black English (BE). They attempted to adjust for population differences by eliminating responses to 16 test items that were incorrect for at least 50% of their subjects. Despite this adjustment, they still found that "... 51% scored more than one standard deviation below the mean for the test's normative data ..." (p. 331). Applying the X^2 procedure to the distribution of adjusted standard scores in each standard deviation category, a non-normal distribution of scores was again obtained ($X^2 = 43.35$, df =

5, $p < 0.000$).

It is important to note however, that both Flipsen (1993) and Washington and Craig (1992) employed larger samples (176 and 105 respectively) than the current study. This highlights the second sampling question: whether the current study had sufficient statistical power. A posthoc power analysis of the current data, using the sample size of 74, 5 degrees of freedom and a type I error rate of 0.05, indicated a power level of approximately 0.70 (Marascuilo & Serlin, 1988) suggesting a 70% likelihood of finding a statistically significant difference, if one existed. Thus, there appeared to be only a mild risk that the current findings were the product of a type II error.

The applicability of the PPVT-R to the current population is further supported by the finding that the number of children failing the screening criterion was not significantly different from expectations based on a normal distribution. The test would then appear to be sufficiently sensitive to identify children in this population with reduced skill. This observation is particularly important given that the specific objective of a screening protocol is identification of those with delays. As with content validity however, the current investigation does not permit examination of the concurrent or predictive validity of the PPVT-R.

Analysis of the basal data also provided support for the use of the PPVT-R. It was only necessary to change direction in testing to ascertain basal levels in just over 1 in 5 cases (compared to more than 9 in 10 cases for the ROWPVT). This suggests a more time-efficient application with the PPVT-R as compared to the ROWPVT. Clinicians thus should be able to use the recommended starting points for PPVT-R testing with this population with much less concern about having to reverse direction in testing.

Of course, with this type of analysis, one must always be aware of potential cohort problems. The two tests were administered to two different groups of children. This is not likely a major concern however because the two sample groups represented contiguous cohorts from the same community. As well, in both cases the entire cohort was tested suggesting no sampling bias.

Overall, the findings suggest that the PPVT-R is a much more suitable instrument for assessing the level of receptive vocabulary skill with small-town Canadian kindergarten children than the ROWPVT. Performance was normally distributed, it appeared to be adequately sensitive as a screening instrument, and administration did not need be modified from standard procedures for maximum efficiency.

Studies such as this one, combined with the results of Flipsen (1993) and Washington and Craig (1992), underscore the continued need for clinicians to be wary of the application of standardised instruments to populations that differ in systematic ways from that used as the standardization sample (McCauley & Swisher, 1984). Given the commonplace use of tests normed on American children by Canadian clinicians, additional studies of this type are strongly recommended.

There are at least three alternatives to conducting studies such as the current one. The first is development of independent tests for Canadian children. Given the size of the Canadian market, it seems unlikely that publishers will be motivated to support such efforts or invest in the marketing of such instruments. A second alternative might be to develop independent Canadian norms for existing instruments. While this would appear to be a less expensive option, little effort appears to have been made in this direction to date. As well, it assumes that the items for the test are appropriate for Canadian children, an assumption requiring independent verification, weakening the cost-effectiveness argument. The third alternative would be to include a proportional number of Canadian children in the normative samples of tests developed in the United States. An informal survey of 10 American tests commonly used by Canadian clinicians (see Appendix) found that such inclusion only occurred twice (Hresko, Reid, & Hammill, 1981; Newcomer & Hammill, 1988). This suggests that test developers are not highly motivated to include Canadian children. It has been argued that including minority groups is less than desirable because the data on the minority group (Canadians here) would be lost in that of the majority (Americans) making it impossible to discern the performance of the two groups (Vaughn-Cooke, 1983; Washington & Craig, 1992). Without the ability to discern the groups, the test user is left to assume that the groups did not differ in their performance when they may have, as was shown by Washington and Craig. In conclusion, further studies such as the current one would appear to be highly warranted.

Acknowledgements

The author gratefully acknowledges the support and encouragement of School District administrative staff during the data collection phase of this project. The assistance of the kindergarten teachers was also much appreciated. Preparation of this paper was supported, in part, by a grant from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health (R01 DC00496).

Please address all correspondence to: Peter Flipsen Jr., MA, S-LP(C), University of Wisconsin-Madison, Waisman Center on Mental Retardation and Human Development, 1500 Highland Ave., Madison, WI, 53705-

2280 USA. Email: flipsen@waisman.wisc.edu

Submitted: March, 1997
Accepted: January, 1998

References

Bailey Jr., D. B., & Wolery, M. (1989). *Assessing Infants and Preschoolers with Handicaps*. Columbus, OH: Merrill.

Baker, R. C., Kummer, A. W., Schultz, J. R., Ho, M., & Gonzalez del Rey, J. (1996). Neurodevelopmental outcome of infants with viral meningitis in the first three months of life. *Clinical Pediatrics*, *35*, 295-301.

Dawson, P. W., Blamey, P. J., Dettman, S. J., Barker, E. J., & Clark, G. M. (1995). A clinical report on receptive vocabulary skills in cochlear implant users. *Ear and Hearing, 16*, 287-294.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test - Revised.* Circle Pines, MN: American Guidance Service.

Eskenazi, B., & Trupin, L. S. (1995). Passive and active maternal smoking during pregnancy, as measured by serum cotinine, and postnatal smoke exposure. II. Effects on neurodevelopment at age 5 years. *American Journal of Epidemiology, 142 (Supplement)*, S19-29.

Flipsen Jr., P. (1993). Use of the ROWPVT with small-town Canadian kindergarten children. *Journal of Speech-Language-Pathology and Audiology*, *17*, 145-148.

Gardner, M. F. (1985). *Receptive One-Word Picture Vocabulary Test.* Novato, California: Academic Therapy Publications.

Hresko, W. P, Reid, D. K., & Hammill, D. D. (1981). *The Test of Early Language Development (TELD)*. Austin, TX: Pro-Ed.

Marascuilo, L. A., & Serlin, R. C. (1988). *Statistical Methods for the Social and Behavioral Sciences.* New York: W. H. Freeman and Company.

McCauley, R. J., & Swisher, L. (1984). Use and misuse of norm-referenced tests in clinical assessment: a hypothetical case. *Journal of Speech and Hearing Disorders*, *49*, 338-348.

Minitab Inc. (1995). Minitab Release 10Xtra [software program]. State College, PA: Minitab Inc.

Newcomer, P., & Hammill, D. D. (1988). *Test of Language Development - 2: Primary (TOLD2-P).* Austin, TX: Pro-Ed.

O'Callaghan, M., Williams, G. M., Anderson, M. J., Bor, W., & Najman, J. M. (1995). Social and biological risk factors for mild and borderline impairment of language comprehension in a cohort of five-year-old children. *Developmental Medicine and Child Neurology*, *37*, 1051-1061.

Statistics Canada (1991). *1991 Census of Population. Area Profile Series: Profile of Census Divisions and Subdivisions in British Columbia (Part B).* Ottawa, ON: Statistics Canada.

Sturner, R. A., Layton, T. L., Evans, A. W., Heller, J. W., Funk, S. G., & Machon, M. W. (1994). Preschool speech and language screening: A review of currently available tests. *American Journal of Speech-Language Pathology, 3(1)*, 25-36.

Triola, M. F. (1980). *Elementary Statistics.* Menlo Park, CA: Benjamin/Cummings Publishing Company.

Vaughn-Cooke, F. B. (1983). Improving language assessment in minority children. *ASHA, 25(9)*, 29-34.

Washington, J. A., & Craig, H. K. (1992). Performances of low-income, African-American preschool and kindergarten children on the Peabody picture vocabulary test-revised. *Language, Speech and Hearing Services in Schools*, *23*, 329-333.

APPENDIX

Standardized Tests Reviewed.

Brown, L., Sherbenou, R. J., & Johnson, S. K. (1990). *Test of Nonverbal Intelligence - Second Edition (TONI-2).* Austin, TX: Pro-Ed.

Carrow-Woolfolk, E. (1985). *Test for Auditory Comprehension of Language - Revised (TACL-R).* Austin, TX: Pro-Ed.

Fensen, L., Dale, P. S., Reznick, S., Thal, D., Bates, E., Hartung, J.P., Pethick, S., & Reilly, J. S. (1993). *MacArthur Communicative Development Inventory (MCDI).* San Diego, CA: Singular.

Gardner, M. F. (1979). *Expressive One-Word Picture Vocabulary Test (EOWPVT).* Novato, VA: Academic Therapy.

Hresko, W. P, Reid, D. K., & Hammill, D. D. (1981). *The Test of Early Language Development (TELD).* Austin, TX: Pro-Ed.

Newcomer, P., & Hammill, D. D. (1988) *Test of Language Development - 2: Primary (TOLD2-P).* Austin, TX: Pro-Ed.

Semel, E., Wiig, E. H., & Secord, W. (1987). *Clinical Evaluation of Language Fundamentals - Revised (CELF-R).* New York: Psychological Corporation.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Stanford-Binet Intelligence Scale - Fourth Edition (S-B).* Chicago, IL: Riverside.

Wallace, G., & Hammill, D. D. (1994). *Comprehensive Receptive and Expressive Vocabulary Test (CREVT).* Austin, TX: Pro-Ed.

Zimmerman, I. L., Steiner, V. G., & Pond, R. V. (1993). *Preschool Language Scale - Third Edition (PLS-3).* New York: Psychological Corporation.