Running Head: READING ALTERNATE ASSESSMENT GENERALIZABILITY TABLES

Alternate Assessment Item and Rater Generalizability for Expressive and Receptive Formats

Gerald Tindal and Paul Yovanoff

University of Oregon

Variables

The purpose of the study was to estimate various sources of error associated with different items on a reading assessment that had been administered using an expressive or receptive communication format. Many students with the most significant disabilities do not express themselves in traditional ways but may need to point, nod, blink, or use any number of alternatives to indicate their response. We consider all of these as receptive communication and studied these contrasting modes of communication on several different reading tasks.

Insert Table 1 about here

Because we were interested in studying reading as a holistic construct in addition to the various tasks noted in Table 1, we also trained judges (master's degree students in a special education program) to make a judgment on each student's total performance.

Insert Table 2 about here

We organized our research by fully specifying all of the variables, some of which were the direct focus of investigation (task, item, format, and rater) and others of which were controlled or served as the context (form, occasion, teacher, and state). An important dimension of generalizability research is specifying whether the variables in the universe of admissible evidence are random or fixed.

Insert Table 3 about here

To allow each student the opportunity to take both the reading tasks in both formats, we needed to cross persons by form. However, to avoid having the items in each task repeated, we then had to create two forms (A and B). Furthermore, to ensure an order effect did not confound the results, we also had to randomly assign either of the two forms to be administered first or second. In the end, we therefore had four measurement conditions in which receptive forms A and B were administered and expressive forms A and B were administered either first or second.

Insert Table 4 about here

While the primary purpose of our research was to study the generalizability of the measures, we completed some preliminary item analyses investigating the classical and Rasch based reliability analyses, the Rasch modeled item functioning, and the equivalence of measurement forms A and B.

Preliminary Analyses

Classical Reliability

Measurement reliability from a classical perspective is very informative with respect to understanding overall consistency of performance across obtained observations. Table 5 provides coefficient alpha estimates per task by format.

Insert Table 5 about here

Rasch Item Analyses. Applying the Rasch partial credit model (Masters, *), analyses indicate that the items generally function well with respect to hypotheses of item difficulty (Table 6). However, in contrast to the extremely high coefficient alpha reliabilities, the Rasch-based reliability estimates reported in Table 7 are considerably lower. Extreme scores inflate classical reliability estimates, e.g. alpha, which is often recognized as an over-estimation of reliability. Rasch-based estimates adjust for extreme scores, considering them as measurement error, providing lower estimates than those obtained classically. The Rasch Partial Credit Model reliability estimates may be considered the lower bound on reliability.

Insert Table 6 about here

Insert Table 7 about here

Equivalence of Form A and B. It is important to establish the equivalence of measurement forms A and B prior to presentation of the G Study results. The data collection design administered randomly either form A or form B to each participant. No participants completed both forms. Table 8 reports the descriptive statistics for each form by expressive and receptive formats. Note, scores are reported as proportion correct. Table 9 reports the MANOVA results testing if the forms are of equivalent difficulty within format. The Forms main effect is significant only for the expressive administration format.

Insert Table 8 about here

Insert Table 9 about here

Generalizability Study and Decision Study Analyses

Our G Study and D Study results focus on the reliability of (a) items within tasks, items within administration format, and (c) rater within administration format. A series of Generalizability studies and Decision studies were designed to estimate the variance associated with item sampling and measurement reliability. Twelve independent studies were completed for each of the six tasks by the two administration formats. For each D Study, reliabilities (Generalizability Coefficients) were estimated for measurement designs with between 3 and 10 items. Tables 10 through 15 report the results for each of the six reading tasks, with results for both the expressive and receptive administration formats. In addition to the studies of generalizability across item

samples per task by format, we completed G and D studies for generalizability across raters. The rater reliability results are reported in Table 16.

Before presenting our results it is helpful to review the purposes of the generalizability analyses. The G study provides the variance component estimates which are then used to estimate error and reliability for each of the D studies. D study estimates include the following statistics.

- $\sigma^2(\tau)$: *variance of universe (true) scores.*
- $\sigma^2(\delta)$: *relative error variance* (difference between observed deviation score and universe deviation score) similar to classical theory error variance.
- $\sigma^2(\Delta)$: *absolute error variance* (1.8, 1.9), difference between observed and universe score, square root is the "absolute SEM providing confidence intervals for universe scores.
- $E\rho^2$: *generalizability coefficient* (ratio of universe score variance to itself plus relative error variance); interpretable as classical theory reliability.
- Φ: *index of dependability* (ratio of universe score variance to itself plus absolute error variance) appropriate when scores are given absolute interpretations as in domain or criterion-referenced measurements.

The interaction component *p x i* indicates the extent to which student relative standing varied from item to item. Larger values indicate greater error and lower reliability.

Signs & Symbols Identification (Task 1). Table 10 reports the G and D study results for the Signs and Symbols Identification task. The error associated with item sampling is reflected in the size of the G Study item variance component estimate and the p x i estimate. Larger values result in lower absolute ($E\rho^2$) and relative (Φ) reliability estimates. From Table 10, it is clear that the item sampling is much less reliable for the expressive format than for the receptive format. For instance, with an item sample of 8, the expressive generalizability coefficient $E\rho^2$ equals .73, while for the receptive format with 8 items $E\rho^2$ equals .85. Even with 10 items, the expressive format is not adequately sampled.

Insert Table 10 about here

Letter Names (Task 2). Table 11 reports G and D Study results for Letter Names. For this task, the item reliability is extremely high for both expressive and receptive formats. Rather 10 items, high reliability could be achieved with as few as three items.

Insert Table 11 about here

Word Reading (Task 3). Table 12 reports G and D Study results for the Word Reading task. The measurement reliability differs slightly depending on format, though item sampling of word reading is quite reliable. For the expressive format, a measurement reliability of .85 is achieved with four items, while with the receptive format five items are necessary.

Insert Table 12 about here

Sentence Reading (Task 4). Table 13 reports the G and D Study results for the Sentence Reading task. This task is reliably sampled for both the expressive and receptive formats. The expressive

format is slightly more reliable, with only 3 item sample necessary for a .85 reliability compared to four items for the receptive format.

Insert Table 13 about here

Passage Reading (Task 5). Table 14 reports G and D Study results for the Passage Reading task. Reliability for this task is very different for the two formats. The expressive format is much more reliable. A reliability of .85 is achieved with as few as four items, while for the receptive format, eight items are necessary for a comparable measurement reliability.

Insert Table 14 about here

Passage Comprehension (Task 6). Table 15 reports the G and D study results for Passage Comprehension. This task appears difficult to sample reliably irrespective of administration format. A reliability of 0.85 is achieved with 10 items for the expressive format, while 10 receptive formats results in a reliability of .83, only.

Insert Table 15 about here

Figure 2 and Figure 3 provide a graphic display of the D study results for expressive and receptive adminstration formats, respectively. The generalizability coefficient estimates for each of the six tasks are displayed, for measurement designs having 3 through 10 items. From these graphs, it is clear that Signs and Symbol Identification, and Passage Comprehension are relatively difficult to sample reliably for the expressive format. When preparing receptive format items, Signs and Symbol Identification, Word Reading, Passage Reading, and Passage Comprehension are difficult to sample reliably.

Rater G and D Study Results

The rater G and D study results are reported in Table 16. The G Study estimates were obtained with each person being rated by each of five raters. The raters were carefully trained to consider performance on each of the six tasks and make an overall rating of 0 (no ability) through 5 (full ability). Ratings were obtained for both admininistration formats. The G study variance components associated with raters are extremely small relative to the person variance. This results in a very high level of measurement reliability. Using only two raters, the generalizability coefficient is .935 for expressive format and .927 for the receptive format. Obviously, there is little need to increase the number of raters. These results are very consistent with other findings reported in the literature (e.g., Brennan, 2000).

Insert Table 16 about here

Insert Figures 2-4 about here

Tables and Figures

Table 1. Number of Items per Task by Format

Task	Format	
	Expressive	Receptive
1. Identify Signs and Symbols	10	10
2. Letter Naming	10	10
3. Word Reading	10	10
4. Sentence Reading	5	5
5. Passage Reading	6	6
6. Comprehension Passage	6	6

Table 2. Rating Scale of Overall Student Performance Aggregated Across All Six Tasks.

0	1	2	3	4	5
Student demonstrates no behavior.	Student demonstrates virtually no comprehension and few symbol, letter, or word skills.	Student demonstrates limited comprehension with some symbol, letter, and word reading skills.	Student demonstrates emerging comprehension with some sentence reading and few passage reading skills.	Student demonstrates basic comprehension with accurate word and sentence reading skills, and some passage reading skills.	Student demonstrates full comprehension, including accurate symbol, letter, word, sentence, and passage reading.

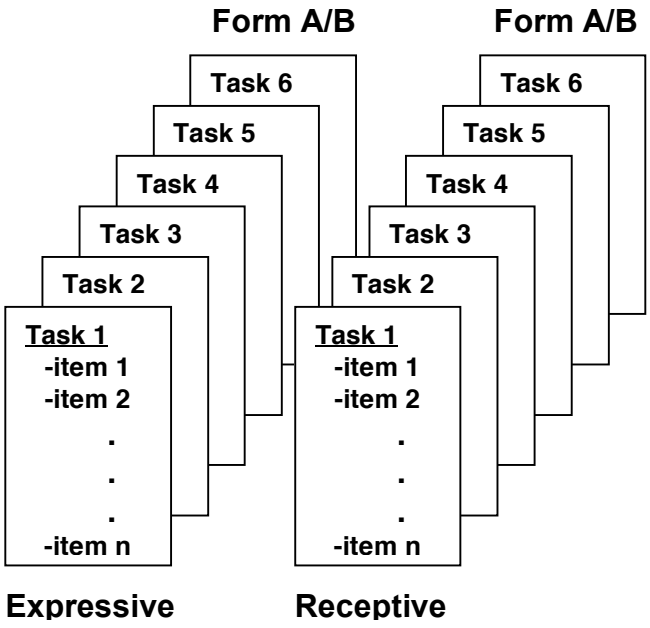

Figure 1. Form, Task, and Item Sampling Structure.

Table 3. Generalizability Study Specifications

Facet	N	Universe of Admissibility	Sampling Notes
persons	n_p=81	*Random* sample of student population	The population is the *object of measurement*, and is not considered a design facet,
task	n_t =6	*Fixed* restricted to specific skills	Fully crossed with persons
item	n_t =variable per task	*Random* sampling from infinite universe of items	Sampling is the conceptualization of item preparation
format	n_t =2	*Fixed* restricted to Expressive and Receptive	Expressive,Receptive Fully crossed with persons and tasks
form	n_f =2	*Random* sample A, B of an infinite set of forms	Multiple 'parallel' forms were created by expert staff item development
occasion	n_o =2	*Random* sample 1, 2 of an infinite set of occassions	
rater	n_r=5	*Random* sample 1, 2, . . ., 5 of an infinite set of raters	All forms rated by each of 5 raters; Rater sampling is the conceptualization for rater training, rubric preparation, etc.
teacher	n_{tch}=66	*Random* sample 1, 2, . . ., 66 of an infinite set of teachers	Sampled within states
state	n_s=9	*Random* sample 1, 2, . . ., 7 of an infinite set of sets	Volunteer recruitment through personal contacts

Table 4. Numbers of Participants Taking Combinations of Forms (A/B) by Format (Receptive/Expressive)

Forms	N	%
rA/eA	20	24.7
rA/eB	19	23.5
rB/eA	26	32.1
rB/eB	16	19.8
Total	81	100.0

Table 5. Classical Task Reliability Estimates (Coefficient Alpha) by Format by Form

	Coefficient Alpha					
	Expressive			Receptive		
		Form			Form	
Task	n items	A	B	n items	A	B
1. Signs/Symbols Identification	10	.87	.85	10	.92	.86
2. Letter Names	10	.96	.96	10	.96	.94
3. Word Reading	10	.95	.96	10	.92	.92
4. Sentence Reading	4	.95	.93	5	.88	.86
5. Passage Reading	5	.88	.96	6	.84	.78
6. Passage Comprehension	6	.88	,89	6	.75	.73

Table 6. Average Rasch Mean-Square Fit Statistics per Task by Type by Form

	Average 'Mean-Square' Fit					
	Expressive			Receptive		
		Form			Form	
Task	n items	A	B	n items	A	B
1. Signs/Symbols Identification	10	0.98	1.00	10	1.16	1.29
2. Letter Names	10	0.85	1.24	10	1.09	0.94
3. Word Reading	10	0.92	0.84	10	1.31	0.93
4. Sentence Reading	4	0.85	0.90	5	0.96	1.00
5. Passage Reading	5	0.88	0.80	6	1.02	0.97
6. Passage Comprehension	6	0.80	0.88	6	1.05	1.00

Note. Rasch fit statistics, e.g. mean square error, ranges between 0.5 and 1.5. Values above 1.5 indicate random error and indicate item level unreliability.

Table 7. Rasch Partial Credit Model Task Reliability by Format by Form

	Rasch-Based Reliability					
	Expressive			Receptive		
		Form			Form	
Task	n items	A	B	n items	A	B
1. Signs/Symbols Identification	10	.60	.38	10	.53	.64
2. Letter Names	10	.47	.67	10	.74	.44
3. Word Reading	10	.77	.83	10	.48	.55
4. Sentence Reading	4	.87	.86	5	.53	.67
5. Passage Reading	5	.94	.95	6	.69	.54
6. Passage Comprehension	6	.70	.67	6	.67	.72

Note. Extreme scores inflate classical reliability estimates, e.g. alpha, which is often recognized as an over-estimation of reliability. Rasch-based estimates adjust for extreme scores, considering them as measurement error, providing lower estimates than those obtained classically. The Rasch Partial Credit Model reliability estimates may be consider the lower bound on reliability.

Table 8. Task Proportion Correct Descriptive Statistics by Format by Form

	Expressive				Receptive			
	A		B		A		B	
Task	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	0.47	0.27	0.28	0.25	0.71	0.35	0.64	0.33
2	0.80	0.28	0.57	0.44	0.77	0.37	0.76	0.35
3	0.33	0.34	0.22	0.34	0.47	0.39	0.53	0.38
4	0.45	0.39	0.31	0.38	0.60	0.41	0.56	0.40
5	0.43	0.39	0.33	0.40	0.41	0.37	0.46	0.34
6	0.17	0.27	0.14	0.27	0.55	0.33	0.39	0.31

Table 9. Task by Form by Format Multivariate (Hotelling's Trace) MANOVA Summary

Effect	F	df1	df2	Eta Square
Expressive				
Task (within)	28.48**	5	75	0.66
Task by Form (within)	2.49*	5	75	0.15
Form (between)	4.13*	1	79	0.05
Receptive				
Task (within)	22.31**	5	75	0.60
Task by Form (within)	3.47**	5	75	0.19
Form (between)	0.17	1	79	0.002

*p <.05, **p <.01

Table 10. Signs & Symbols Identification—Item Expressive and Receptive Format Random Effects G/D Study Variance Component and Reliability Estimates

		Expressive							
		D Study							
Effect	G Study	3	4	5	6	7	8	9	10
person (p)	.049	.049	.049	.049	.049	.049	.049	.049	.049
items (i)	.012	.004	.003	.002	.002	.001	.001	.001	.001
p x i	.015	.048	.036	.029	.024	.021	.018	.016	.014
	$\sigma^2(\tau)$	.049	.049	.049	.049	.049	.049	.049	.049
	$\sigma^2(\delta)$	.048	.036	.029	.024	.020	.018	.016	.015
	$\sigma^2(\Delta)$	.052	.039	.031	.026	.022	.019	.017	.015
	$E\rho^2$	.502	.573	.627	.668	.702	.729	.751	.770
	Φ	.482	.554	.608	.650	.684	.713	.736	.756
		Receptive							
		D Study							
Effect	G Study	3	4	5	6	7	8	9	10
person (p)	.094	.094	.094	.094	.094	.094	.094	.094	.094
items (i)	.004	.001	.001	.000	.000	.000	.000	.000	.000
pi	.127	.042	.031	.025	.021	.018	.015	.014	.012
	$\sigma^2(\tau)$	.094	.094	.094	.094	.094	.094	.094	.094
	$\sigma^2(\delta)$	.042	.031	.025	.021	.018	.015	.014	.012
	$\sigma^2(\Delta)$	.043	.032	.026	.021	.018	.016	.014	.013
	$E\rho^2$	.691	.749	.788	.817	.838	.856	.870	.881
	Φ	.683	.742	.782	.812	.834	.852	.866	.878

Table 11. Letter Names—Item Expressive and Receptive Format Random Effects G/D Study Variance Component and Reliability Estimates

		Expressive							
		D Study							
Effect	G Study	3	4	5	6	7	8	9	10
person (p)	.149	.149	.149	.149	.149	.149	.149	.149	.149
items (i)	.001	.000	.000	.000	.000	.000	.000	.000	.000
pi	.066	.022	.017	.013	.011	.009	.008	.007	.006
	$\sigma^2(\tau)$	.149	.149	.149	.149	.149	.149	.149	.149
	$\sigma^2(\delta)$	.022	.016	.013	.011	.009	.008	.007	.006
	$\sigma^2(\Delta)$	.022	.017	.013	.011	.009	.008	.007	.006
	$E\rho^2$	.870	.899	.917	.930	.939	.947	.952	.957
	Φ	.868	.898	.916	.929	.939	.946	.952	.956
		Receptive							
		D Study							
Effect	G Study	3	4	5	6	7	8	9	10
person (p)	.120	.120	.120	.120	.120	.120	.120	.120	.120
items (i)	.000	.000	.000	.000	.000	.000	.000	.000	.000
pi	.059	.019	.015	.012	.010	.008	.007	.006	.006
	$\sigma^2(\tau)$	.120	.120	.120	.120	.120	.120	.120	.120
	$\sigma^2(\delta)$	.020	.015	.011	.010	.008	.007	.006	.006
	$\sigma^2(\Delta)$	.020	.015	.011	.010	.008	.007	.006	.006
	$E\rho^2$	.858	.890	.910	.924	.934	.942	.948	.953
	Φ	.858	.889	.909	.923	.933	.941	.947	.952

Table 12. Word Reading—Item Expressive and Receptive Format Random Effects G/D Study Variance Component and Reliability Estimates

		Expressive							
		D Study							
Effect	G Study	3	4	5	6	7	8	9	10
person (p)	.097	.097	.097	.097	.097	.097	.097	.097	.097
items (i)	.001	.001	.000	.000	.000	.000	.000	.000	.000
pi	.068	.023	.017	.014	.011	.010	.009	.008	.007
	$\sigma^2(\tau)$	.097	.097	.097	.097	.097	.097	.097	.097
	$\sigma^2(\delta)$	.023	.017	.014	.011	.010	.009	.008	.007
	$\sigma^2(\Delta)$	.023	.017	.014	.011	.010	.009	.008	.007
	$E\rho^2$	.810	.850	.876	.895	.910	.920	.927	.934
	Φ	.807	.848	.875	.893	.907	.918	.926	.933
		Receptive							
		D Study							
Effect	G Study	3	4	5	6	7	8	9	10
person (p)	.131	.131	.131	.131	.131	.131	.131	.131	.131
items (i)	.004	.001	.001	.001	.001	.001	.001	.000	.000
pi	.117	.039	.029	.023	.019	.017	.015	.013	.012
	$\sigma^2(\tau)$	.131	.131	.131	.131	.131	.131	.131	.131
	$\sigma^2(\delta)$	.039	.029	.023	.019	.017	.015	.013	.012
	$\sigma^2(\Delta)$	.040	.030	.024	.020	.017	.015	.013	.012
	$E\rho^2$	.771	.818	.849	.871	.887	.899	.910	.918
	Φ	.764	.812	.844	.866	.883	.896	.907	.915

Table 13. Sentence Reading—Item Expressive and Receptive Format Random Effects G/D Study Variance Component and Reliability Estimates

		Expressive							
		D Study							
Effect	G Study	3	4	5	6	7	8	9	10
person (p)	.125	.125	.125	.125	.125	.125	.125	.125	.125
items (i)	.000	.000	.000	.000	.000	.000	.000	.000	.000
pi	.065	.022	.016	.013	.011	.009	.008	.007	.007
	$\sigma^2(\tau)$	.125	.125	.125	.125	.125	.125	.125	.125
	$\sigma^2(\delta)$	.022	.016	.013	.011	.009	.008	.007	.007
	$\sigma^2(\Delta)$	.022	.016	.013	.011	.009	.008	.007	.007
	$E\rho^2$	.851	.884	.905	.920	.930	.938	.945	.950
	Φ	.851	.884	.905	.920	.930	.938	.945	.950
		Receptive							
		D Study							
Effect	G Study	3	4	5	6	7	8	9	10
person (p)	.142	.142	.142	.142	.142	.142	.142	.142	.142
items (i)	.001	.000	.000	.000	.000	.000	.000	.000	.000
pi	.103	.034	.026	.021	.017	.015	.013	.011	.010
	$\sigma^2(\tau)$	.142	.142	.142	.142	.142	.142	.142	.142
	$\sigma^2(\delta)$	.034	.026	.021	.017	.015	.013	.011	.010
	$\sigma^2(\Delta)$	.035	.026	.021	.017	.015	.013	.012	.010
	$E\rho^2$	.806	.847	.874	.892	.906	.917	.926	.932
	Φ	.804	.846	.873	.892	.906	.916	.925	.932

Table 14. Passage Reading—Item Expressive and Receptive Format Random Effects G/D Study Variance Component and Reliability Estimates

		Expressive							
		D Study							
Effect	G Study	3	4	5	6	7	8	9	10
person (p)	.087	.087	.087	.087	.087	.087	.087	.087	.087
items (i)	.002	.001	.001	.000	.000	.000	.000	.000	.000
pi	.060	.020	.015	.012	.010	.009	.008	.007	.006
	$\sigma^2(\tau)$	.087	.087	.087	.087	.087	.087	.087	.087
	$\sigma^2(\delta)$	.020	.015	.012	.010	.009	.008	.007	.006
	$\sigma^2(\Delta)$	.021	.016	.013	.010	.009	.008	.007	.006
	$E\rho^2$	.812	.852	.878	.897	.910	.920	.929	.935
	Φ	.807	.848	.875	.893	.907	.918	.926	.933
		Receptive							
		D Study							
Effect	G Study	3	4	5	6	7	8	9	10
person (p)	.102	.102	.102	.102	.102	.102	.102	.102	.102
items (i)	.001	.000	.000	.000	.000	.000	.000	.000	.000
pi	.145	.048	.036	.029	.024	.021	.018	.016	.015
	$\sigma^2(\tau)$	.102	.102	.102	.102	.102	.102	.102	.102
	$\sigma^2(\delta)$	.048	.036	.029	.024	.021	.018	.016	.015
	$\sigma^2(\Delta)$	.049	.037	.029	.024	.021	.018	.016	.015
	$E\rho^2$	.678	.737	.779	.808	.831	.849	.863	.875
	Φ	.676	.736	.777	.807	.830	.848	.862	.874

Table 15. Passage Comprehension—Item Expressive and Receptive Format Random Effects G/D Study Variance Component and Reliability Estimates

		Expressive							
		D Study							
Effect	G Study	3	4	5	6	7	8	9	10
person (p)	.036	.036	.036	.036	.036	.036	.036	.036	.036
items (i)	.003	.001	.001	.001	.001	.001	.000	.000	.000
pi	.066	.022	.017	.013	.011	.009	.008	.007	.007
	$\sigma^2(\tau)$	.036	.036	.036	.036	.036	.036	.036	.036
	$\sigma^2(\delta)$	.022	.017	.013	.011	.009	.008	.007	.007
	$\sigma^2(\Delta)$	.023	.017	.014	.012	.010	.010	.010	.007
	$E\rho^2$	.621	.686	.732	.766	.793	.814	.831	.845
	Φ	.609	.675	.722	.757	.784	.806	.824	.838
		Receptive							
		D Study							
Effect	G Study	3	4	5	6	7	8	9	10
person (p)	.081	.081	.081	.081	.081	.081	.081	.081	.081
items (i)	.008	.003	.002	.002	.001	.001	.001	.001	.000
pi	.162	.054	.041	.032	.027	.023	.020	.018	.016
	$\sigma^2(\tau)$	.081	.081	.081	.081	.081	.081	.081	.081
	$\sigma^2(\delta)$	.054	.041	.032	.027	.023	.020	.018	.016
	$\sigma^2(\Delta)$	.057	.043	.034	.028	.024	.021	.019	.017
	$E\rho^2$	.600	.667	.714	.750	.778	.800	.818	.833
	Φ	.588	.656	.704	.741	.769	.792	.811	.826

Table 16. Rater Expressive and Receptive Format Random Effects G/D Study Variance Component and Reliability Estimates

Effect	G Study	2	3	4	5	6	7
person (p)	1.586	1.586	1.586	1.586	1.586	1.586	1.586
rater (r)	.023	.012	.008	.006	.005	.004	.003
pr	.211	.106	.071	.053	.042	.035	.030
	$\sigma^2(\tau)$	1.586	1.586	1.586	1.586	1.586	1.586
	$\sigma^2(\delta)$	.106	.071	.053	.042	.035	.030
	$\sigma^2(\Delta)$	.117	.078	.059	.047	.039	.033
	$E\rho^2$	.937	.957	.968	.974	.978	.981
	Φ	.931	.953	.964	.971	.976	.979
Effect	G Study	2	3	4	5	6	7
person (p)	1.960	1.960	1.960	1.960	1.960	1.960	1.960
rater (r)	.037	.019	.012	.009	.007	.006	.005
pr	.317	.159	.106	.079	.063	.053	.045
	$\sigma^2(\tau)$	1.960	1.960	1.960	1.960	1.960	1.960
	$\sigma^2(\delta)$	.159	.106	.079	.063	.053	.045
	$\sigma^2(\Delta)$	.177	.118	.090	.071	.059	.051
	$E\rho^2$	.925	.949	.961	.969	.974	.977
	Φ	.917	.943	.957	.965	.971	.975

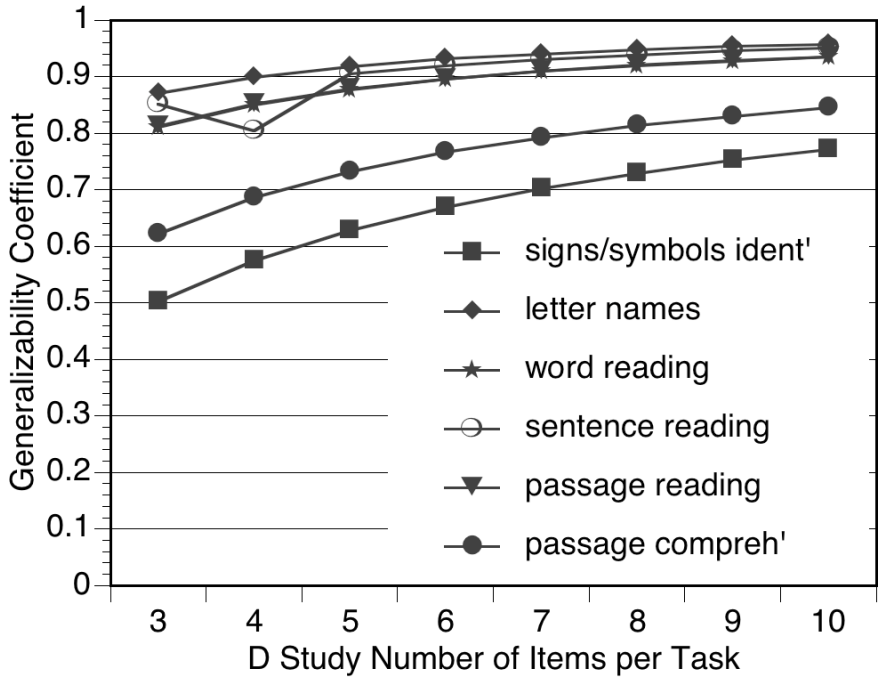

Figure 2. Reading Task Expressive Administration Format D Study Generalizability Coefficient Estimates.

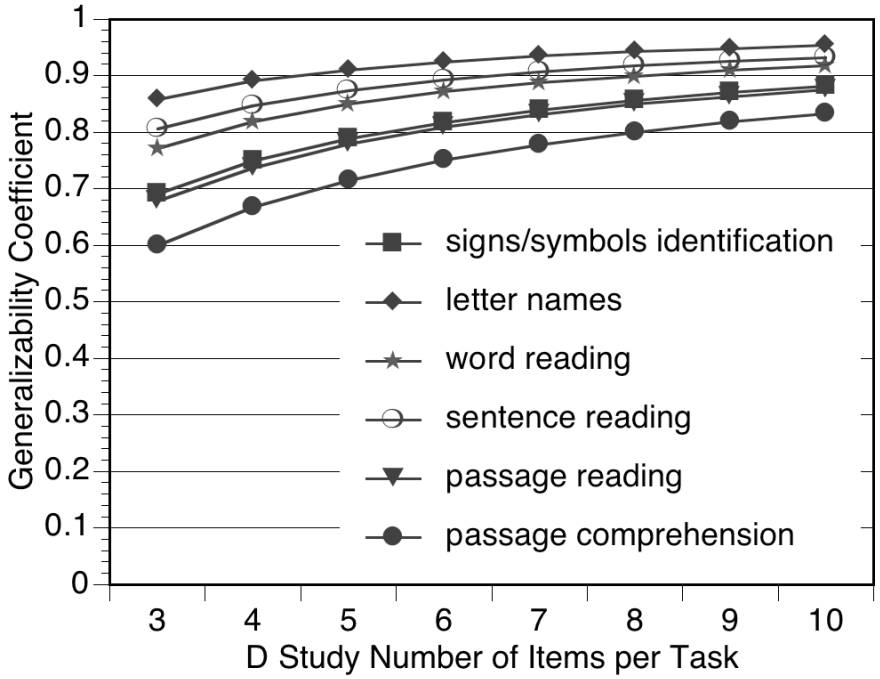

Figure 3. Reading Task Receptive Administration Format D Study Generalizability Coefficient Estimates

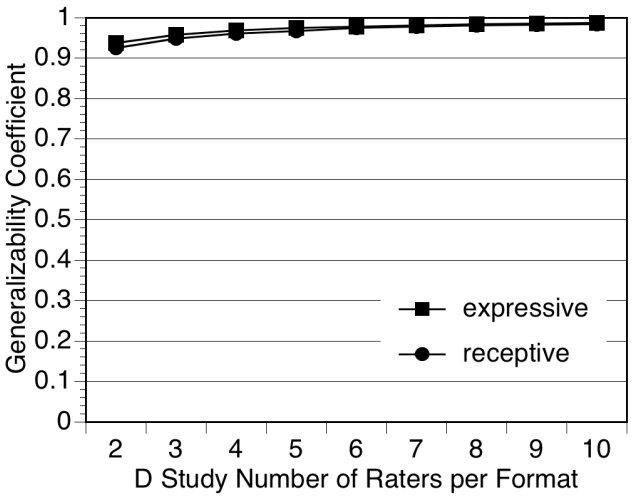

Figure 4. Rater D Study Generalizability Coefficient Estimates