

F

Technical Report on ASQ



This appendix contains information relating to the development and psychometric studies completed on the *Ages & Stages Questionnaires*¹ at the Center on Human Development, University of Oregon, Eugene. First, the development of the ASQ system, including item selection, readability, item analyses, and psychometric studies, is addressed. Second, information about the demographic characteristics of the samples used to study the ASQ system is given. Third, reliability analyses, including internal consistency, alphas, correlation coefficients, and standard error of measurement, are provided. Validity studies are provided fourth, including a description of the determination of the cutoff points, relative operating characteristic (ROC) analysis, concurrent validity, sensitivity, specificity, and overreferral and underreferral rates. The final section of this appendix contains information about risk and nonrisk samples, including performance and effects of risk status.

DEVELOPMENT OF THE AGES & STAGES QUESTIONNAIRES

Item Selection

ASQ items were developed using a variety of sources, including standardized developmental tests, nonstandardized tests focused on early development, textbooks, and other literature containing information about early developmental milestones. Using these sources, the following criteria were used to develop items:

1. Skills were selected that could be easily observed or elicited by parents.
2. Skills were selected that were highly likely to occur in a home setting.

Once skills were selected, items were written in familiar, nonjargon wording not to exceed a sixth-grade reading level, illustrations were provided when possible, and concrete examples were provided as appropriate.

Appreciation is extended to Craig Leve for data analysis and to Roland Good for data consultation.

¹The 60 month questionnaire was in the final stages of development when these studies were conducted. Results of validity studies to date are included.

The five items that make up each developmental area (i.e., communication, gross motor, fine motor, problem solving, and personal-social) were chosen to represent as closely as possible the developmental quotient (DQ) range of 75–100. This range was chosen for two reasons. First, many standardized tests use 1.5–2.0 standard deviations below the mean as the lower end of the typical developmental range; therefore, it was reasoned that any child who was generally unable to perform items at a developmental quotient of 75 should be referred for further assessment. For example, Knobloch, Stevens, and Malone (1980), in assigning diagnostic categories for the Revised Gesell and Amatruda Developmental and Neurological Examination, designated a developmental quotient of 75 or less as the cutoff for atypical test scores. Second, it was reasoned that items above a developmental quotient of 100 would primarily identify children whose development was clearly within typical limits and, thus, the inclusion of such items would add little in attempting to identify children whose development was suspect. Limiting the questionnaires to the developmental range of 75–100 also assisted in maintaining the brevity of the questionnaires.²

In order to determine the developmental quotient for each item, the following formula was used:

$$\frac{\text{age equivalence}}{\text{test interval of item}} \times 100 = \text{DQ}$$

The age equivalence was obtained from the source(s) of the item such as the Gesell (Knobloch et al., 1980), the Bayley Scales of Infant Development (Bayley, 1969), and *Developmental Resources: Behavioral Sequences for Assessment and Program Planning* (Cohen & Gross, 1979). When sources varied, a developmental range was used. Table 1 contains the developmental quotient and age equivalent for each item by area for each of the 19 questionnaires. An examination of Table 1 indicates that each area has two items with developmental quotients of approximately 75 (range: 69–77), two items with developmental quotients of approximately 85 (range: 80–92), and two items with developmental quotients of approximately 100 (range: 91–120).

Reading Level

The *Ages & Stages Questionnaires* were designed to be used with a range of parents (e.g., varying income and educational levels); therefore, the reading level ranges between the fourth- and sixth-grade levels, and illustrations were added when possible to clarify items. To ascertain the reading level of the questionnaires, the MinNesota Interactive Readability Approximation Program (MECCA, n.d.) was used, which performed several common readability tests including the Dale-Chall, Raygor, and Fry indexes. Readability levels of the questionnaires ranged from fourth to sixth grade on these three measures, depending on the questionnaire interval and index, with a mean 5.8 grade level across age intervals and indexes.

²One “validity check” item with an approximate developmental quotient of 125–150 was also included in each area on the initial version of the ASQ.

Table 1. Age equivalent and developmental quotient of items by area for each questionnaire

Questionnaire items	Communication		Gross motor		Fine motor		Problem solving		Personal-social	
	Age	DQ	Age	DQ	Age	DQ	Age	DQ	Age	DQ
4 Month										
1	12w ^{a,b}	75	12w	75	12w	75	8-12w	75	12w	75
2	12-16w ^c	75-100	12w	75	12w	75	12w	75	12w	75
3	12-16w	75-100	8-12w	75	12w	75	12w	75	12w	75
4	16w	100	16w	100	16w	100	16w	100	16w	100
5	16w	100	16w	100	16w	100	16w	100	16w	100
6	16w	100	16w	100	16w	100	16w	100	16w	100
6 Month										
1	16w	62	20w	77	20w	77	20w	77	20w	77
2	20w	77	20w	77	20w	77	20w	77	20w	77
3	24w	92	24w	92	24w	92	24w	92	24w	92
4	24w	92	24w	92	24w	92	24w	92	24w	92
5	28w	107	28w	107	28w	107	28w	107	28w	107
6	28w	107	28w	107	28w	107	28w	107	28w	107
8 Month										
1	24w	69	24w	69	24w	69	24w	69	24w	69
2	24w	69	24w	69	24w	69	24w	69	24w	69
3	28w	80	28w	80	28w	80	28w	80	28w	80
4	28w	80	28w	80	28w	80	28w	80	28w	80
5	32w	91	28-32w	80-91	32w	91	32w	91	32w	91
6	32w	91	32w	91	36w	103	32w	91	32w	91
10 Month										
1	28w	70	28w	70	28w	70	28w	70	28w	70
2	28w	70	28-32w	70-80	32w	80	32w	80	32w	80
3	32w	80	32w	80	36w	90	32w	80	32w	80
4	40w	100	40w	100	40w	100	40w	100	40w	100
5	44w	110	40w	100	40w	100	40w	100	40w	100
6	44w	110	44w	110	44w	110	44w	110	44w	110

^aNumbers were rounded to the nearest whole numbers.

^bw = weeks.

^cRanges are presented when the age and developmental quotient (DQ) of an item differed according to developmental sources.

^dm = months.

(continued)

Table 1. (continued)

Questionnaire items	Communication		Gross motor		Fine motor		Problem solving		Personal-social	
	Age	DQ	Age	DQ	Age	DQ	Age	DQ	Age	DQ
12 Month										
1	40w	77	40w	77	40w	77	40w	77	40w	77
2	44w	85	40w	77	40w	77	40w	77	40w	77
3	44w	85	44w	85	44w	85	44w	85	44w	85
4	48w	92	44w	85	48w	92	44w	85	44w	85
5	52w	100	48w	92	48w	92	48w	92	48w	92
6	52w	100	52w	100	52w	100	52w	100	52w	100
14 Month										
1	44w	80	44w	79	48w	86	44w	79	44w	79
2	52w	93	48w	86	48w	86	48w	86	48w	86
3	52w	93	52w	93	52w	93	52w	93	52w	93
4	52w	93	52-56w	93-100	56w	100	52w	93	56-60w	100-107
5	52w	93	52w	93	60w	107	52w	93	48-60w	86-107
6	56w	100	56w	100	60w	107	56w	100	52-56w	93-100
16 Month										
1	52w	75	52w	75	52w	75	52w	75	15m ^d	93.75
2	52w	75	56w	81.25	52w	75	52w	75	12-15m	75-94
3	52w	75	52w	75	56w	81.25	56w	81	52w	75
4	56w	81.25	56w	81.25	15m	93.75	56w	81	52w	75
5	15m	93.75	15m	93.75	15m	93.75	15m	93.75	12m	94
6	56w	81.25	15m	93.75	18m	112.5	15m	93.75	15m	94
18 Month										
1	56w	74	52w	68	52w	68	56w	74	52w	68
2	56w	74	56w	74	56w	74	56w	74	52w	68
3	65w	85	65w	85	65w	85	65w	85	65w	85
4	65w	85	15m	83	15m	83	65w	85	15m	83
5	78w	108	18m	100	18m	100	78w	102	78w	102
6	91w	126	18m	100	18m	100	78w	102	78w	102

Table 1. (continued)

Questionnaire items	Communication		Gross motor		Fine motor		Problem solving		Personal-social	
	Age	DQ	Age	DQ	Age	DQ	Age	DQ	Age	DQ
33 Month										
1	24m	73	21m	64	24m	73	24m	73	24m	73
2	24m	73	24m	73	30m	91	24m	73	24m	73
3	30m	91	24m	73	30m	91	24m	73	30m	91
4	30m	91	30m	91	30m	91	30m	91	30m	91
5	36m	109	30m	91	30m	91	30m	91	30m	91
6	36m	109	36m	109	30m	91	30m	91	36m	109
36 Month										
1	24m	67	24m	67	24m	67	24m	67	24m	67
2	24m	67	24m	67	24m	67	24m	67	24m	67
3	30m	83	30m	83	30m	83	30m	83	30m	83
4	30m	83	30m	83	30m	83	30m	83	30m	83
5	36m	100	36m	100	36m	100	36m	100	36m	100
6	36m	100	36m	100	36m	100	36m	100	36m	100
42 Month										
1	30m	71	30m	71	30m	71	30m	71	30m	71
2	30m	71	30m	71	30m	71	30m	71	30m	71
3	36m	86	36m	86	36m	86	36m	86	36m	86
4	36m	86	36m	86	36m	86	36m	86	36m	86
5	36-48m	86-114	36-57m	86-135	36-48m	86-114	36-57m	86-135	36-48m	86-114
6	36-49m	86-117	45-60m	107-117	42m	100	42m	100	31-49m	74-117

48 Month												
1	40-72m	88-150	36-57m	75-119	36-48m	75-100	36-48m	75-100	36-48m	75-100	36-48m	75-100
2	54-60m	113-125	45-60m	94-125	36-57m	75-108	36-57m	75-119	48-60m	48-60m	48-60m	100-125
3	30-60m	63-125	36-48m	75-100	48-60m	100-125	36-53m	75-111	36-54m	75-113	48m	75-113
4	48-60m	100-125	35m	73	48m	100	41-53m	85-111	48m	100	48m	100
5	36-48m	75-100	36-48m	75-100	48m	100	42m	88	42-60m	88-125	42-60m	88-125
6	36-49m	75-102	36-72m	75-150	48m	100	36-44m	75-92	31-49m	65-102	31-49m	65-102
54 Month												
1	36-60m	66-111	35m	65	48-60m	88-111	41-53m	76-98	36-54m	66-100	36-54m	66-100
2	48-60m	88-111	36-48m	66-88	48m	88	42m	77	48m	88	48m	88
3	36-49m	66-91	36-48m	66-88	48m	88	36-44m	81-82	42-60m	77-111	42-60m	77-111
4	36-48m	66-88	36-57m	66-106	45m	83	36-57m	82-106	36-48m	66-88	36-48m	66-88
5	48m	88	36-72m	66-133	54m	100	54m	100	48-60m	88-111	48-60m	88-111
6	48-59m	88-109	54-60m	100-111	48-57m	88-106	53-60m	98-111	51-66m	94-122	51-66m	94-122
60 Month												
1	36-48m	60-80	36-48m	60-80	45m	75	36-57m	60-75	36-48m	60-80	36-48m	60-80
2	48m	80	36-57m	60-95	54m	90	41-53m	68-88	36-54m	60-90	36-54m	60-90
3	48-59m	80-98	36-72m	60-120	48-57m	80-95	54m	90	48-60m	80-100	48-60m	80-100
4	54-60m	90-100	54-60m	90-100	48-60m	80-100	60m	100	51-66m	85-110	51-66m	85-110
5	54-60m	90-100	60m	100	48-60m	80-100	53-60m	88-100	51-66m	85-110	51-66m	85-110
6	54-60m	90-100	60-66m	100-110	54-66m	90-110	60m	100	48-62m	80-103	48-62m	80-103



Revisions of the Ages & Stages Questionnaires

First Revision In response to validity and utility data gathered on the questionnaires (e.g., Brinker, Franzier, Lancelot, & Norman, 1989), the questionnaires were revised in 1991. Five types of changes were made. First, a number of items were reworded to clarify meaning. These modifications were made based on feedback from project staff, interventionists, parents, nurses, and pediatricians using the questionnaires in clinic and research environments. In most cases, the modifications entailed minimal word changes. For example, "reach for a toy" was changed to "try to get a toy"; "couch or adult chair" was changed to "furniture"; "being able to stop" was changed to "stopping"; and "stack" was changed to "stack on top of." In a few cases, examples were added or modified. For example, "Does your baby play ball with you by either rolling or throwing the ball to you?" was changed to "Does your baby either

roll or throw a ball back to you so that you can return it to him?" For some items, examples were expanded to include the use of familiar and more available household objects to facilitate completion of the questionnaires by parents from diverse households. For example, "toy" and "four objects like blocks or cars" were substituted for "block" in several instances. In other cases, illustrations were added or modified to clarify the intent of the item. For example, one illustration was modified so that it was clear that the infant was using her hands for assistance rather than sitting unassisted with a straight back.

Second, modifications of a more extensive nature were made. In some cases, an item that was difficult to interpret was eliminated and replaced with another item. In all cases, the substituted items appeared on an ASQ at the previous or next interval. For example, on the 20 month questionnaire, an item in the fine motor area was eliminated and replaced with an item from the fine motor area of the 24 month questionnaire. The number and type of modifications for each questionnaire can be found in Table 2.

The third change made to revise the questionnaires was the elimination of items with a developmental quotient of 125–150. On the initial version of the questionnaires, each behavior area included one item with a developmental range of 125–150. These additional five items per questionnaire were added to provide information on parents' reported tendency to overestimate their children's developmental status (cf. Gradel, Thompson, & Sheehan, 1981; Hunt & Paraskevopoulos, 1980).

To examine how parents scored the 30 validity check items, the percentage of parents who scored these items *yes* was calculated. Table 3 indicates by

Table 2. Number of items per questionnaire with minor and major revisions

Questionnaire interval ^a	Minor wording revisions	Major revisions or substitutions
4	20	1
8	28	1
12	28	0
16	26	1
20	25	3
24	26	2
30	23	1
36	23	0

^aThe questionnaires targeting 48 and 60 months were not developed when this analysis was completed.

area the percentage of parents who selected *yes* for these items. The percentage of *yes* responses to these 30 items ranged from a low of 4% to a high of 81% (with a mean of 21%). This analysis does not support the notion that parents overestimate their children’s developmental achievements. Rather, the analysis suggests that there is wide variation in the developmental achievements of children. Further support for the accuracy of parental judgments of children’s developmental skills is found in an earlier study of the questionnaires (Bricker & Squires, 1989) in which interrater agreement on the questionnaires for 112 parents and several trained examiners exceeded 90%. Finally, on a number of occasions, trained examiners corroborated parental reports by observing infants and children exhibiting behaviors far above their general developmental level. For these reasons, items exceeding a DQ of 125 were eliminated from the revised questionnaires.

A fourth change was the ordering of items within each area according to their developmental order. As mentioned previously, items appeared in random developmental order within each area on the original version of the questionnaires. The five items in each area (e.g., communication, fine motor) were rearranged, beginning with the lowest age item and moving to the highest age item.

A fifth modification was the addition of the 6, 18, and 48 month questionnaires. The 6 and 18 month questionnaires were constructed by taking developmentally appropriate items from the adjacent questionnaires and adding items when necessary. The 48 month questionnaire was developed by examining a variety of tests and other developmental resources and constructing test items. The same criteria for the development of the previous questionnaires were applied to items for the 48 month questionnaire.

Table 3. The percentage of parents scoring *yes* on the 30 items with developmental quotients over 125 by area and questionnaire interval

Questionnaire (by month)	n	Percent of <i>yes</i> responses				
		Communication	Gross motor	Fine motor	Problem solving	Personal-social
4	625	81	27	35	24	24
8	594	14	17	40	7	12
12	562	8	40	10	7	35
16	547	15	13	10	18	10
20	472	4	39	24	25	7
24	503	13	14	44	13	14

Second Revision A second revision of the questionnaires was completed in June of 1994. Revisions were minor, and little adjustment of the items was undertaken. This revision included three types of modifications: name changes, minor modification of items, and format changes.

First, a new name, *Ages & Stages Questionnaires*,³ was adopted to be more appealing to parents and professionals. Second, minor wording modifications were made to increase the clarity of items. For example, qualifying words such as "generally" or "usually" were eliminated. Finally, the questionnaire format was modified to be more user friendly.

Additional Intervals From 1997 to 1998, additional intervals were completed at 10, 14, 22, 27, 33, 42, 54, and 60 months. These intervals were added to make the ASQ series more comprehensive and to avoid screening children whose ages fall outside the validity "window." Validity and reliability studies were begun on the 60 month (5 year) ASQ. The remaining additional intervals (including 6 and 8 months) have not been studied; cutoff points were determined by estimating developmental quotients using age equivalencies. All items appear on questionnaires that have been empirically studied, however.

Ages & Stages Questionnaires: Social-Emotional Also in 1997, with the passage of the amendments to the Individuals with Disabilities Education Act (IDEA), came a call for early detection of social or emotional problems in young children. The *Ages & Stages Questionnaires: Social-Emotional* is a screening tool, meant to be used in conjunction with the ASQ, to identify the need for further social and emotional behavior assessment in children from 3 to 60 months of age. Eight questionnaires are available, in either English or Spanish, that address seven behavioral areas: self-regulation, compliance, communication, adaptive functioning, autonomy, affect, and interaction with people. An accompanying *User's Guide* is also available to assist professionals in the effective use of the ASQ:SE questionnaires.

DEMOGRAPHIC CHARACTERISTICS OF SAMPLES

The data reported in this section include those from questionnaires completed by parents of children between 4 and 36 months old, primarily for the Infant/Child Monitoring Project in Oregon. Additional programs, such as the Easter Seals' Watch-Me-Grow Program originating in Youngstown, Ohio, and the University of Hawaii Department of Pediatrics, have supplied data from their projects on various parameters of the questionnaires. Although gender of child data were collected in all cases, other demographic variables such as ethnicity, family income, and education level of parents were not always collected because of concerns about family privacy. We have included in our analyses children with missing demographic information.

The total number of children in the original sample was 2,008.⁴ Of these, 53% ($n = 1,068$) were male, and 47% ($n = 940$) were female. Ethnicity of the sample is displayed in Table 4. Level of income of participating families is contained in Table 5. Occupational status of mother and father are contained in Tables 6 and 7, respectively. Level of education of mother and father can be

³The questionnaires were originally known as the Infant/Child Monitoring Questionnaires.

⁴Not all children or parents are included in all studies of the questionnaires. Most validity and reliability studies included only a subsample of children and parents. An additional 320 families participated in studies of the 48 and 60 month questionnaires.

Table 4. Ethnicity of sample

Ethnicity	<i>n</i>	Percent
Caucasian	835	64.9
African American	174	13.5
Asian/Pacific Islander	5	0.4
Native American/Alaskan	188	14.6
Latino/Hispanic	52	4.0
Biracial	33	2.6
Total	1,287	100.00

Table 5. Family income level

Income (in dollars)	<i>n</i>	Percent
Less than 5,000	125	12.70
5,001–10,000	131	13.30
10,001–15,000	120	12.20
15,001–20,000	148	15.10
20,001–25,000	141	14.30
More than 25,000	296	30.10
More than 40,000	23	2.30
Total	984	100.00

Table 6. Occupational status of mother

Mother's occupation	<i>n</i>	Percent
Student	45	4.70
Unskilled laborer	495	51.90
Semiskilled laborer	33	3.50
Skilled laborer	17	1.80
Clerical	187	19.60
Administrator	77	8.10
Manager	90	9.40
Executive	9	0.90
Total	953	100.00

Table 7. Occupational status of father

Father's occupation	<i>n</i>	Percent
Student	26	3.20
Unskilled laborer	80	10.00
Semiskilled laborer	159	19.90
Skilled laborer	113	14.10
Clerical	156	19.50
Administrator	160	20.00
Manager	75	9.40
Executive	32	4.00
Total	801	100.00

Table 8. Level of mother's education

Education level	<i>n</i>	Percent
Less than seventh grade	29	1.60
Junior high	87	4.90
Partial high school	313	17.80
High school graduate	510	29.00
Partial college	469	26.60
College graduate	233	13.20
Professional	119	6.80
Total	1,760	100.00

found in Tables 8 and 9, respectively. In some cases, when no additional clarification was given on the demographic form, it was assumed that information collected on the "primary caregiver" was the mother and on the "secondary caregiver," the father.

Between 1980 and 1988, infants with medical risk factors discharged from the neonatal intensive care unit (NICU) of a regional medical facility constituted the primary research sample included in studies by the Center on Human Development. Beginning in 1988, two additional groups of children were studied: 1) those from so-called families at environmental risk as a result of economic or social conditions; and 2) young children from nonrisk, or normative, households.

Medical Risk Sample

Subjects from the medical risk sample were infants who remained at least 3 days in a Level IV NICU needing medical care for a specific problem (e.g., respiratory distress, prematurity, low birth weight).

Environmental Risk Sample

Subjects from the environmental risk sample were infants and young children from families who met one or more of the following criteria: 1) extreme poverty (according to family income level, as defined by federal guidelines, 100% poverty level); 2) maternal age of 19 years or younger at the time of the infant's birth; 3) maternal education less than twelfth grade; and/or 4) parents who had experienced involvement with children's protective services for abuse and/or neglect of their children. Infants with medical risk factors were excluded from this sample.

Table 9. Level of father's education

Education level	<i>n</i>	Percent
Less than seventh grade	8	1.00
Junior high	10	1.20
Partial high school	70	8.40
High school graduate	239	28.60
Partial college	244	29.20
College graduate	173	20.70
Professional	91	10.90
Total	835	100.00

Table 10. Children by risk status

	Male		Female	
	<i>n</i>	Percent	<i>n</i>	Percent
Risk	869	81.37	751	79.89
Nonrisk	199	18.63	189	20.11
Total	1,068	100.00	940	100.00

Table 11. Ethnicity by risk status

Ethnicity	Nonrisk		Risk	
	<i>n</i>	Percent	<i>n</i>	Percent
Caucasian	108	89.30	727	62.40
African American	3	2.50	171	14.70
Asian/Pacific Islander	1	.80	4	0.30
Native American/Alaskan	4	3.30	184	15.80
Latino/Hispanic	5	4.10	47	4.00
Biracial	0	0.00	33	2.80
Total	121	100.00	1,166	100.00

Table 12. Family income by risk status

Income (in dollars)	Nonrisk		Risk	
	<i>n</i>	Percent	<i>n</i>	Percent
Less than 5,000	12	3.10	113	18.80
5,001–10,000	15	3.90	116	19.30
10,001–15,000	25	6.50	95	15.90
15,001–20,000	55	14.30	93	15.50
20,001–25,000	74	19.30	67	11.20
More than 25,000	189	47.70	113	18.80
More than 40,000	20	5.20	3	0.50
Total	384	100.00	600	100.00

Normative Sample

Subjects from the normative sample were infants and young children who met the following criteria: 1) no previous history of developmental or serious health problems as reported by parents, 2) birth at full term (greater than 37 weeks'), and 3) no assignment to a NICU.

Children are categorized by risk status in Table 10. Because of the overlap among risk factors, the two risk categories—medical and environmental—were combined into one risk category for analyses. Ethnicity by sample, income level by sample, and education level by sample are contained in Tables 11, 12, and 13. The total numbers of questionnaires completed at each age interval are listed in Table 14. The total numbers of questionnaires completed for each subject are shown in Table 15. The maximum number of questionnaires that could be completed for each subject was eight (i.e., at 4, 8, 12, 16, 20, 24, 30, and 36 months); however, for most subjects ($n = 950$) only one questionnaire was completed. For 198 subjects, six consecutive questionnaires were completed between 4 and 24 months. For 62, or 2% of subjects, all eight questionnaires between 4 and 36 months were completed.

Table 13. Parents' educational level

Level	Mother				Father			
	Nonrisk		Risk		Nonrisk		Risk	
	<i>n</i>	Percent	<i>n</i>	Percent	<i>n</i>	Percent	<i>n</i>	Percent
Less than seventh grade	0	0.00	29	2.10	0	0.00	8	1.70
Junior high	0	0.00	87	6.30	1	0.30	9	2.00
Partial high school	8	2.10	305	22.20	8	2.10	62	13.50
High school graduate	83	21.40	427	31.20	99	26.40	140	30.50
Partial college	127	32.80	342	24.90	100	26.70	144	31.30
College graduate	111	28.70	122	8.90	97	25.90	76	16.50
Professional	58	15.00	61	4.40	70	18.70	21	4.50
Total	387	100.00	1,373	100.00	375	100.00	460	100.00

Table 14. Total questionnaires by age interval

Age interval (in months)	<i>n</i>
4	1,500
8	1,405
12	1,185
16	1,057
20	930
24	898
30	609
36	535

Table 15. Number of questionnaires completed on each subject

Number of questionnaires	<i>n</i>	Percent
1	950	33
2	587	21
3	438	15
4	345	12
5	175	6
6	198	7
7	106	4
8	62	2
Total	2,861	100

Note: Questionnaires for which there was no demographic information are included in this analysis.

RELIABILITY

Reliability studies on the ASQ system are described here. Internal consistency analyses, including correlational analyses and Cronbach's coefficient alpha (Cronbach, 1951), are included. Test-retest reliability, interobserver reliability, and standard error of measurement are discussed.

Internal Consistency

The internal consistency of the questionnaires was addressed by examining the relationship between developmental area and overall scores. Cor-

Table 16. Correlations between area and overall score on the 10 questionnaires

Age interval (in months)	<i>n</i>	Communication	Gross motor	Fine motor	Problem solving	Personal-social
4	869	.71	.70	.81	.81	.79
8	768	.72	.76	.79	.79	.79
12	617	.75	.70	.77	.78	.83
16	502	.75	.54	.76	.75	.73
20	443	.75	.70	.66	.77	.71
24	393	.69	.63	.74	.76	.76
30	305	.76	.69	.73	.83	.69
36	248	.77	.77	.78	.83	.73
48	336	.73	.69	.82	.66	.75
60	125	.44	.58	.55	.55	.48

Note: All correlations are significant at $p < .0001$.

relational analyses and Cronbach's coefficient alpha (Cronbach, 1951) were calculated.

Pearson product moment correlation coefficients were calculated for each area score with an overall score for individual questionnaires as shown in Table 16. The overall score was obtained by adding the five developmental area scores. Correlations ranged from .70 to .81 for the 4 month ASQ, from .72 to .79 for the 8 month, from .70 to .83 for the 12 month, from .54 to .76 for the 16 month, from .66 to .77 for the 20 month, from .63 to .76 for the 24 month, from .69 to .83 for the 30 month, from .73 to .83 for the 36 month, from .66 to .82 for the 48 month, and from .44 to .58 for the 60 month. All correlations were significant at $p < .0001$.

Pearson product moment coefficient correlations between developmental area and overall scores across questionnaires are contained in Table 17. Again, all correlations were significant at $p < .0001$.

Cronbach's coefficient alpha was calculated for area scores on individual questionnaires. For the communication area, alphas ranged from .63 at 4 months to .75 at 24 months. For the gross motor area, alphas ranged from .53 at 4 months to .87 at 12 and 16 months. The fine motor area had a coefficient alpha range of .49 at 20 months to .79 at 8 months. For the problem solving area, alphas ranged from .52 at 20 months to .75 at 8 months. Finally, for the personal-social area, alphas ranged from .52 at 16 months to .68 at 12 months. Table 18 contains the standardized alphas by area and age interval.

These ranges of alphas are the result of several factors. First, the varying developmental quotients of the items that compose each area are unlikely to result in a high alpha. As a result, a child performing below criterion in a particular area will not have a static score across items within that area. Second,

Table 17. Correlations between area scores collapsing across questionnaires

Area	Communication	Gross motor	Fine motor	Problem solving	Personal-social
Communication					
Gross motor	0.46				
Fine motor	0.46	0.49			
Problem solving	0.64	0.52	0.51		
Personal-social	0.48	0.51	0.39	0.59	
Overall	0.77	0.77	0.78	0.83	0.73

N = 4,145.

Note: All correlations are significant at $p < .0001$.

Table 18. Standardized alphas by area and age interval

Age interval (in months)	<i>n</i>	Communication	Gross motor	Fine motor	Problem solving	Personal- social
4	848	0.63	0.53	0.66	0.68	0.61
8	743	0.65	0.76	0.79	0.75	0.66
12	591	0.65	0.87	0.68	0.69	0.68
16	490	0.68	0.87	0.65	0.57	0.52
20	416	0.72	0.76	0.49	0.52	0.53
24	367	0.75	0.80	0.58	0.57	0.58
30	285	0.74	0.78	0.70	0.61	0.56
36	231	0.69	0.76	0.72	0.66	0.55
48	336	0.79	0.84	0.86	0.85	0.86
60	125	0.79	0.75	0.76	0.77	0.77

the error variance in a measure is increased when a statistic uses individual items rather than an aggregated total for a calculation. In this case, the reported alphas use the actual items from each area, thus the possible error variance is increased. In contrast, the Pearson product moment correlation coefficients are a product of area totals when error variance is reduced simply because individual items are combined.

Test-Retest Reliability

Test-retest reliability was determined by comparing the results of two questionnaires completed by parents in a 2-week time period. To assess the test-retest reliability of the questionnaires, parents who brought their infants to the center for a standardized assessment completed a second identical questionnaire immediately before the standardized assessment was administered. The two questionnaires completed by parents were then compared for agreement. Test-retest reliability, measured as percentage agreement between classifications based on the questionnaires completed by 175 parents at 2-week intervals, was 94%. The standard error of measurement was .10.

Interobserver Reliability

Interobserver reliability was examined by comparing infants' classifications based on questionnaires completed by parents with the classifications based on questionnaires completed by examiners immediately after the standardized assessments, as described previously. Interobserver reliability, measured as percentage agreement between classifications based on the questionnaires completed by 112 parents and those completed by 2 examiners, was 94%. The standard error of measurement was .12. Seventy-four protocols were eliminated from this analysis because one or more test areas had two or more uncompleted items. This occurred because the professional examiner had little opportunity to observe children engaged in certain activities (e.g., eating, pretend play, adaptive skills). For example, on the 20 month ASQ, the item, "Does your child copy activities you do, such as wipe up a spill, vacuum, shave, or comb hair?" was difficult for examiners to observe and, therefore, was not scored. There is no reason to believe that any particular bias was operating in the elimination of infants from this analysis.

The reliability of the questionnaires has been studied by examining the internal consistency, test-retest reliability, and interobserver reliability of the questionnaires. Internal consistency analyses indicated strong relationships across items and within areas on the questionnaires. The questionnaires also

achieved substantial test–retest and interobserver reliability. Parents' evaluations of their children using the questionnaires were consistent over time. In addition, professional examiners' agreement with parental evaluations of children using the questionnaires was consistently high.

VALIDITY

Studies of the validity of the ASQ system are described here. Relative operating characteristic (ROC) analyses; determination of screening cutoff points; and studies on concurrent validity including sensitivity, specificity, overreferral, underreferral, and positive predictive value are included.

Determination of Screening Cutoff Points

Sample The first question to be addressed was the determination of the sample to be used for calculating referral cutoff points.⁵ Including both the risk and nonrisk samples is more representative and would likely provide more accurate cutoff points. The method used to test this question was an analytic technique called relative (or receiver) operating characteristic. The ROC, based on statistical decision theory, was developed in the context of electronic signal detection (Peterson, Birdsall, & Fox, 1954) and has been used in a variety of disciplines, including human perception and decision making (Green & Swets, 1966). The ROC provides estimates of the probabilities of decision outcomes by revealing the trading relationship between the true positive, true negative, false positive, and false negative probabilities that can be attained by shifting the decision criteria (i.e., cutoff points).

For this analysis, the ROC was employed to provide a single value measure of accuracy, which is reported as the area of a curve. This value represents the area of the entire graph that lies beneath the curve and can vary between .50 (when no discrimination exists) to 1.0 for perfect discrimination (when the curve follows the left and upper axes, such that the true positive proportion is 1.0 for all values of the false positive proportion). The application of the ROC to the ASQ system was to establish whether cutoff points based on the risk group, the nonrisk group, or a combined group were most accurate.

Curves were generated based on cutoff points derived from the means and standard deviations for the risk, nonrisk, and combined groups, and the areas of each curve were compared for each of the first eight questionnaires. Three cutoff points were used in the calculation of the area under the curve for each group at each questionnaire age interval: 1) 1 standard deviation below the mean, 2) 1½ standard deviations below the mean, and 3) 2 standard deviations below the mean. Each of the three points on the curve were then determined by plotting the true positive probability against the false positive probability for each cutoff point. For example, for the 12 month questionnaire, three separate ROC curves were generated, representing the risk, nonrisk, and combined groups. The ROC curve produced for the combined group using these cutoff points (i.e., 2, 1½, and 1 standard deviation) is shown in Figure 1.

To look at the differences across groups (risk, nonrisk, or combined) more effectively, the proportion of area reported for each group for each question-

⁵Cutoff points refer to the score on the questionnaires below which the infant/child is identified for further testing.

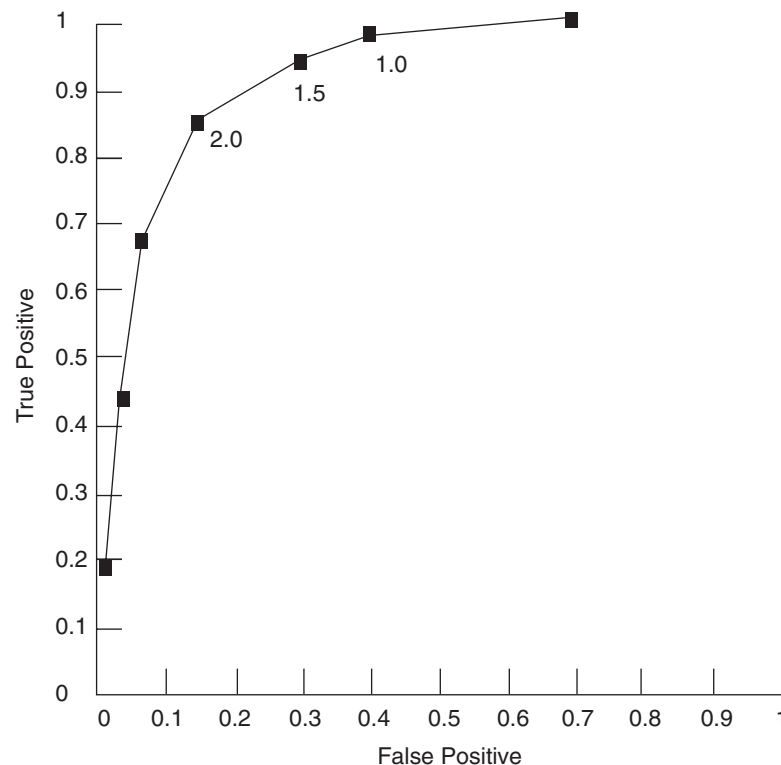


Figure 1. ROC curve generated for 12 month questionnaire combined group using 1, 1½, and 2 standard deviations.

naire was combined, and the mean area for each group was generated. The following range of areas and mean areas were found: 1) for the risk group, areas ranged from .737 to .90, with a mean of .812; 2) for the nonrisk group, areas ranged from .734 to .881, with a mean of .80; and 3) for the combined risk and nonrisk group, areas ranged from .750 to .90, with a mean of .825.

It was determined, based on the reported range of areas and mean areas by group, that points derived by using means and standard deviations from the combined risk and nonrisk group provided the most accurate cutoff points. This decision also has practical implications because agencies responsible for screening often do not know the risk status of the population to be screened. Adopting the combined risk and nonrisk referral cutoffs appeared to be most appropriate.

Cutoff Points Given that the sample used to establish cutoff points included a risk and nonrisk population, or combined group, a related question arose: Which specific cutoff point (1, 1½, or 2 standard deviations below the mean) should be used to maximize the conditional probabilities of true positive and false positive proportions? This question was addressed in two ways. First, a matrix showing the conditional probabilities that exist at each cutoff point was designed. Second, a ROC curve that graphically displayed the probabilities at each cutoff point was generated.

A matrix for each questionnaire interval was designed using the cutoff points at 2, 1½, and 1 standard deviation from the mean. This matrix included several computations that were generated using a simple contingency table.

Table 19. Cutoff points by standard deviation units and accompanying conditional probabilities for the 12 month questionnaire

Standard deviation unit(s) below the mean	Developmental area	Cutoff	Sensitivity	Specificity	True positive	False positive	Overreferral	Underreferral
2	Communication	1.54	.85	.86	.85	.13	.12	.02
	Gross motor	1.80						
	Fine motor	2.87						
	Problem solving Personal-social	2.51 1.98						
1½	Communication	2.21	.95	.71	.95	.28	.25	.006
	Gross motor	2.57						
	Fine motor	3.39						
	Problem solving Personal-social	3.10 2.62						
1	Communication	2.88	.97	.59	.97	.41	.36	.003
	Gross motor	3.33						
	Fine motor	3.90						
	Problem solving Personal-social	3.68 3.26						

		Standardized assessment		Total
		<i>Fail</i>	<i>Pass</i>	
4 month ASQ	<i>Risk</i>	25	23	48
	<i>Developing typically</i>	24	120	144
Total		49	143	192

Sensitivity	Specificity	False positive	True positive	False negative	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
51.02%	83.92%	16.08%	51.02%	48.98%	75.52%	12.50%	11.98%	25.00%	52.08%

		Standardized assessment		Total
		<i>Fail</i>	<i>Pass</i>	
8 month ASQ	<i>Risk</i>	28	20	48
	<i>Developing typically</i>	8	151	159
Total		36	171	207

Sensitivity	Specificity	False positive	True positive	False negative	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
77.78%	88.30%	11.70%	77.78%	22.22%	86.47%	3.46%	9.66%	23.19%	58.33%

		Standardized assessment		Total
		<i>Fail</i>	<i>Pass</i>	
12 month ASQ	<i>Risk</i>	33	39	72
	<i>Developing typically</i>	6	249	255
Total		39	288	327

Sensitivity	Specificity	False positive	True positive	False negative	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
84.62%	86.46%	13.54%	84.62%	15.38%	86.24%	1.83%	11.93%	24.33%	45.83%

		Standardized assessment		Total
		<i>Fail</i>	<i>Pass</i>	
16 month ASQ	<i>Risk</i>	16	31	47
	<i>Developing typically</i>	6	135	141
Total		22	166	188

Sensitivity	Specificity	False positive	True positive	False negative	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
72.73%	81.33%	18.67%	72.73%	27.27%	80.32%	3.19%	16.49%	25.00%	34.04%

(continued)

Figure 2. Cross tabulation of agreement between combined standardized assessments and ASQ by age interval. For the 4, 8, 12, 16, and 20 month comparisons, the standardized assessment instruments used were the Bayley Scales of Infant Development (Bayley, 1969) and the Revised Gesell and Armatruda Developmental and Neurological Examination (Knobloch et al., 1980). For the 24 month comparison, these instruments and the Stanford-Binet Intelligence Scale (Thorndike et al., 1985) were used. The 30 month comparison used all of the above-mentioned instruments and the McCarthy Scales of Children's Abilities (McCarthy, 1972). The 36 month comparison used only the Gesell, the Stanford-Binet Intelligence Scale, and the McCarthy Scales of Children's Abilities. The 48 month comparison used only the McCarthy Scales of Children's Abilities. The Battelle Developmental Inventory (Newborg, Stock, Wnek, Guidubaldi, & Svinicki, 1984) was used at 60 months.

Figure 2. (continued)

		Standardized assessment		Total
		<i>Fail</i>	<i>Pass</i>	
20 month ASQ	<i>Risk</i>	17	13	30
	<i>Developing typically</i>	9	119	128
Total		26	132	158

Sensitivity	Specificity	False positive	True positive	False negative	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
65.38%	90.15%	9.58%	65.38%	34.62%	86.70%	5.70%	8.23%	18.99%	56.67%

		Standardized assessment		Total
		<i>Fail</i>	<i>Pass</i>	
24 month ASQ	<i>Risk</i>	12	37	49
	<i>Developing typically</i>	3	174	177
Total		15	211	226

Sensitivity	Specificity	False positive	True positive	False negative	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
80.00%	82.46%	17.54%	80.00%	20.00%	82.30%	1.33%	16.37%	21.68%	24.49%

		Standardized assessment		Total
		<i>Fail</i>	<i>Pass</i>	
30 month ASQ	<i>Risk</i>	9	19	28
	<i>Developing typically</i>	3	113	116
Total		12	132	144

Sensitivity	Specificity	False positive	True positive	False negative	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
75.00%	85.61%	14.39%	75.00%	25.00%	84.72%	2.08%	13.19%	19.44%	32.14%

		Standardized assessment		Total
		<i>Fail</i>	<i>Pass</i>	
36 month ASQ	<i>Risk</i>	9	5	14
	<i>Developing typically</i>	1	53	54
Total		10	58	68

Sensitivity	Specificity	False positive	True positive	False negative	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
90.00%	91.53%	8.47%	90.00%	10.00%	91.30%	1.45%	7.25%	20.29%	64.29%

(continued)

Figure 2. (continued)

48 month ASQ		Standardized assessment		Total
		<i>Fail</i>	<i>Pass</i>	
	<i>Risk</i>	7	14	21
	<i>Developing typically</i>	2	80	82
Total		9	94	103

Sensitivity	Specificity	False positive	True positive	False negative	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
77.78%	85.11%	14.89%	77.78%	22.22%	84.67%	1.94%	13.59%	20.38%	33.33%

60 month ASQ		Standardized assessment		Total
		<i>Fail</i>	<i>Pass</i>	
	<i>Risk</i>	5	0	5
	<i>Developing typically</i>	1	24	25
Total		6	24	30

Sensitivity	Specificity	False positive	True positive	False negative	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
83.33%	100.00%	0.00%	83.33%	16.67%	96.67%	3.33%	0.00%	16.67%	100.00%

Cutoff points for developmental areas were included, along with conditional probabilities that were computed based on a contingency table. A sample matrix for the 12 month questionnaire can be found in Table 19 with the following conditional probabilities: 1) sensitivity, 2) specificity, 3) true positive proportion, 4) false positive proportion, 5) overreferral, and 6) underreferral. (For definitions and computational formulas, see Chapter 6.) Cross tabulations of agreement between combined standardized assessments and the *Ages & Stages Questionnaires* by age interval can be found in Figure 2. As expected, when the cutoff becomes less conservative (i.e., 1½ or 1 standard deviation from the mean), the overreferral rate increases as the underreferral rate decreases.

Further demonstration of the trading relationship of true positive and false positive proportions as the cutoff point is adjusted is provided with the graphic representation of the ROC curve in Figure 1. It can be seen from Table 19 how the cutoff points differ in their placement on the curve. The 2 standard deviations cutoff point, although not perfect, appeared the most balanced cutoff point in terms of the true positive and false positive proportions.

For every questionnaire interval, the process of evaluation described previously was undertaken, and conditional probabilities across questionnaires were computed to arrive at a mean figure for each. These probabilities are reported in Table 20. The same trend, as described for the 12 month questionnaire, is evident in Table 20 for other questionnaire age intervals. Specifically, the sensitivity, or true positive proportion, could be maximized as the cutoff point was raised but at the expense of specificity, false positive proportions, overreferral, and underreferral.

Table 20. Mean conditional probabilities across questionnaire age intervals by cutoff point

Cutoff point (in standard deviations below the mean)	Sensitivity	Specificity	True positive	False positive	Over-referral	Under-referral
2	.75	.86	.75	.14	.12	.04
1½	.82	.74	.82	.26	.23	.03
1	.91	.60	.91	.40	.35	.01

Based on the analyses conducted using the contingency table conditional probabilities and the ROC curve, a referral cutoff point of 2 standard deviations below the mean across questionnaire intervals was recommended.⁶ Because of the serial nature of the questionnaires, a child with delayed development should be identified at a test interval, even if the child is underreferred at one or two test intervals. Inherent errors will always exist when using screening tools because of the lack of an infallible criterion for measuring effectiveness, the uneven nature of early human development, and the balancing of the conditional probabilities used to determine the test's accuracy.

Concurrent Validity Concurrent validity was measured by comparing the classifications of the child's performance based on the parent-completed questionnaire with the classification of the child's performance on a professionally administered standardized test given within 29 days. The Revised Gesell and Amatruda Developmental and Neurological Examination (Knobloch et al., 1980) and the Bayley Scales of Infant Development (Bayley, 1969) were used for infants up to 30 months of age; the Stanford-Binet Intelligence Scale (Thorndike, Hagen, & Sattler, 1985) and the McCarthy Scales of Children's Abilities (McCarthy, 1972) were used for children 3–4 years old. The Battelle Developmental Inventory (BDI) (Newborg et al., 1984) was used at 60 months (5 years) of age. The child's performance on the standardized test was designated as "identified" if the child's scaled score was equal to or less than 1½ standard deviations below the mean on any scale or subscale. This scaled score (76 on the Bayley Scales of Infant Development, the McCarthy Scales of Children's Abilities, the Stanford-Binet Intelligence Scale, and the BDI; and 75 on the Gesell) was chosen because it was believed that a child scoring at or below this point was suspect for developmental delay and should be seen for further diagnostic assessment. In addition, a 1½-standard deviation delay on a standardized test meets eligibility criteria established by many states for entrance into early intervention programs (Brown & Brown, 1993).

To examine agreement between the questionnaire and the standardized measure, the child's classifications on a developmental test and a questionnaire were compared. A child was considered "identified" when his or her score fell below the cutoff point set at 2 standard deviations below the mean. One of the following four outcomes was possible:

1. Both tests classified the child as developing typically or not "identified."
2. Both tests classified the child as "identified."
3. The standardized measure indicated the child as developing typically, and the questionnaire as "identified."

⁶Agencies with sufficient resources to assess additional children can raise the cutoff points to 1½ or 1 standard deviation.

Ages & Stages Questionnaires		Standardized assessment		Total
		<i>Fail</i>	<i>Pass</i>	
	<i>Risk</i>	161	201	362
	<i>Developing typically</i>	63	1219	1282
Total		224	1420	1644

Sensitivity	Specificity	False positive	True positive	False negative	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
71.88%	85.84%	14.15%	71.88%	28.12%	83.94%	3.83%	12.22%	22.02%	44.47%

Figure 3. Cross tabulation of agreement between standardized assessments and the ASQ system across age intervals 4–48 months. The standardized assessment instruments used were the Bayley Scales of Infant Development (Bayley, 1969), the Gesell (Knobloch et al., 1980), the Stanford-Binet Intelligence Scale (Thorndike et al., 1985), and the McCarthy Scales of Children's Abilities (McCarthy, 1972).

- The questionnaire classified the child as developing typically, and the standardized measure as "identified."

Agreement between standardized assessments and the ASQ system across questionnaires is contained in Figure 3. Contingency tables containing agreement between combined standardized assessments and the ASQ system by age interval can be found in Figure 2. Contingency tables containing agreement between individual standardized tests and the ASQ system by age interval can be found in Figure 4.

Validity with Children with Disabilities A study of the current validity of the questionnaires with a subsample of children with disabilities was undertaken. Children in this study ranged from 4 to 36 months old and were enrolled in state-funded early intervention programs for children with mild to severe disabilities. These children had received a multidisciplinary assessment and were found by professional evaluators to meet state eligibility guidelines for receiving publicly supported early intervention services. Of the 46 children whose parents completed questionnaires, 44, or 96%, were identified as in need of further assessment by the questionnaires (i.e., scored below the established cutoff points).

General Validity The validity of the ASQ system has been evaluated extensively. The concurrent validity of the questionnaires as reported in percent agreement between questionnaires and standardized assessments ranged from 76% for the 4 month ASQ to 91% for the 36 month ASQ, with 84% overall agreement. Sensitivity ranged from 51% for the 4 month ASQ to 90% for the 36 month ASQ, with 72% overall agreement. Specificity of the questionnaires ranged from 81% for the 16 month ASQ to 92% for the 36 month ASQ, with 86% overall agreement. Specificity, or the ability of the ASQ system to correctly identify typically developing children, remained high across questionnaire intervals and standardized assessments. Sensitivity, or the ability to detect delayed development, was lower, averaging 72%. In a separate analysis, however, the ability of the questionnaires to identify children with established developmental delays was high (96%).

RISK AND NONRISK GROUP COMPARISONS

Although it was determined that the combined risk and nonrisk groups provided the best sample for determining referral cutoff points, an analysis of dif-

4 Month ASQ

Bayley Scales of Infant Development				Revised Gesell Developmental Examination			
	Fail	Pass		Fail	Pass		
4 month ASQ	0	2	2	Risk	25	21	46
				<i>Developing typically</i>	24	104	128
Total	0	18	18		49	125	174

Assessment	Sensitivity	Specificity	False positive rate	True positive rate	False negative rate	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
BAYLEY	—	88.89%	11.11%	—	—	88.89%	0.00%	11.11%	11.11%	0.00%
GESELL	51.02%	83.20%	16.80%	51.02%	48.98%	74.14%	13.79%	12.07%	26.44%	54.35%

8 Month ASQ

Bayley Scales of Infant Development				Revised Gesell Developmental Examination			
	Fail	Pass		Fail	Pass		
8 month ASQ	3	5	8	Risk	25	15	40
				<i>Developing typically</i>	6	124	130
Total	5	32	37		31	139	170

Assessment	Sensitivity	Specificity	False positive rate	True positive rate	False negative rate	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
BAYLEY	60.00%	84.38%	15.63%	60.00%	40.00%	81.08%	5.41%	13.51%	21.62%	37.50%
GESELL	80.65%	89.21%	10.79%	80.65%	19.35%	87.65%	3.53%	8.82%	23.53%	62.50%

(continued)

Figure 4. Cross tabulation of agreement between the ASQ system and standardized assessment by age interval.

Figure 4. (continued)

12 Month ASQ

		Bayley Scales of Infant Development			Revised Gesell Developmental Examination						
		Fail	Pass								
12 month ASQ	Risk	7	16	23	12 month ASQ						
	Developing typically	3	139	142	Developing typically						
Total		10	155	165	Total						
Assessment	Sensitivity	Specificity	False positive rate	True positive rate	False negative rate	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value	
BAYLEY	70.00%	90.00%	10.00%	70.00%	30.00%	88.48%	1.81%	9.69%	14.00%	30.43%	
GESELL	89.66%	82.71%	17.29%	89.66%	10.34%	83.95%	1.85%	14.20%	30.25%	53.06%	

16 Month ASQ

		Bayley Scales of Infant Development			Revised Gesell Developmental Examination						
		Fail	Pass								
16 month ASQ	Risk	7	13	20	16 month ASQ						
	Developing typically	1	55	56	Developing typically						
Total		8	68	76	Total						
Assessment	Sensitivity	Specificity	False positive rate	True positive rate	False negative rate	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value	
BAYLEY	87.50%	80.88%	19.12%	87.50%	12.50%	81.58%	1.32%	17.11%	26.32%	35.00%	
GESELL	64.29%	81.63%	18.37%	64.29%	35.71%	79.46%	4.46%	16.07%	24.11%	33.33%	

20 Month ASQ

Bayley Scales of Infant Development				Revised Gesell Developmental Examination						
	Fail	Pass		Fail	Pass					
20 month ASQ	Risk	11	6	17	20 month ASQ	Risk	6	7	13	
	Developing typically	4	51	55		Developing typically	5	68	73	
Total	15	57	72	Total	11	75	86			
Assessment	Sensitivity	Specificity	False positive rate	True positive rate	False negative rate	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
BAYLEY	73.33%	90.38%	9.62%	73.33%	26.67%	86.57%	5.97%	7.46%	23.88%	68.75%
GESELL	54.55%	90.67%	9.33%	54.55%	45.45%	86.05%	5.81%	8.14%	15.12%	46.15%

24 Month ASQ

Bayley Scales of Infant Development				Revised Gesell Developmental Examination						
	Fail	Pass		Fail	Pass					
24 month ASQ	Risk	7	29	36	24 month ASQ	Risk	5	7	12	
	Developing typically	1	111	112		Developing typically	2	61	63	
Total	8	140	148	Total	7	68	75			
Assessment	Sensitivity	Specificity	False positive rate	True positive rate	False negative rate	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
BAYLEY	87.50%	79.28%	20.71%	87.50%	12.50%	79.72%	0.67%	19.59%	24.32%	19.44%
GESELL	71.43%	89.71%	10.29%	71.43%	28.57%	88.00%	2.67%	9.33%	16.00%	41.67%

(continued)

Figure 4. (continued)

24 Month ASQ

		Stanford-Binet Intelligence Scale		
		Fail	Pass	
24 month ASQ	Risk	0	1	1
	Developing typically	0	2	2
Total		0	3	3

Assessment	Sensitivity	Specificity	False positive rate	True positive rate	False negative rate	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
STANFORD-BINET	—	66.67%	33.33%	—	—	66.67%	0.00%	33.33%	33.33%	0.00%

30 Month ASQ

		Bayley Scales of Infant Development		Revised Gesell Developmental Examination				
		Fail	Pass					
30 month ASQ	Risk	7	16	23				
	Developing typically	3	98	101				
Total		10	114	124				
					Fail	Pass		
	30 month ASQ				Risk	2	1	3
					Developing typically	0	5	5
	Total	2	6	8				

McCarthy Scales of Children's Abilities

		Fail	Pass
30 month ASQ	Risk	0	0
	Developing typically	0	4
Total		0	4

Assessment	Sensitivity	Specificity	False positive rate	True positive rate	False negative rate	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
BAYLEY	70.00%	85.96%	14.03%	70.00%	30.00%	84.67%	2.41%	12.90%	18.54%	30.43%
GESELL	100.00%	83.33%	16.67%	100.00%	0.00%	87.50%	0.00%	12.50%	37.50%	66.67%
MCCARTHY	—	100.00%	0.00%	—	—	100.00%	0.00%	0.00%	0.00%	—

(continued)

Figure 4. (continued)

36 Month ASQ

		McCarthy Scales of Children's Abilities		Revised Gesell Developmental Examination						
		Fail	Pass							
36 month ASQ	Risk	0	0	36 month ASQ						
	Developing typically	0	5	Developing typically						
Total		0	5	5	5	10	21	31		
Assessment	Sensitivity	Specificity	False positive rate	True positive rate	False negative rate	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
McCARTHY	—	100.00%	0.00%	—	—	100.00%	0.00%	0.00%	0.00%	—
GESELL	90.00%	90.48%	9.52%	90.00%	10.00%	90.32%	3.23%	6.45%	35.48%	81.82%

36 Month ASQ

		Stanford-Binet Intelligence Scale									
		Fail	Pass								
36 month ASQ	Risk	0	3								
	Developing typically	0	30								
Total		0	33	33							
Assessment	Sensitivity	Specificity	False positive rate	True positive rate	False negative rate	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value	
STANFORD-BINET	—	90.91%	9.09%	—	—	90.91%	0.00%	9.09%	9.09%	0.00%	

48 Month ASQ

McCarthy Scales of Children's Abilities

	Fail	Pass
48 month ASQ		
Risk	7	14
Developing typically	2	80
Total	9	94
		103

Assessment	Sensitivity	Specificity	False positive rate	True positive rate	False negative rate	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
McCARTHY	77.78%	85.11%	14.89%	77.78%	22.22%	84.67%	1.94%	13.59%	20.38%	33.33%

60 Month ASQ

Battelle Developmental Inventory

	Fail	Pass
60 month ASQ		
Risk	5	0
Developing typically	1	24
Total	6	24
		30

Assessment	Sensitivity	Specificity	False positive rate	True positive rate	False negative rate	Percent agreement	Under-referred	Over-referred	Percent referred	Positive predictive value
BATTELLE DEVELOPMENTAL INVENTORY	83.33%	100.00%	0.00%	83.33%	16.67%	96.67%	3.33%	0.00%	16.67%	100.00%

ferences between these groups provided interesting information regarding comparative group performances. Item analyses and analyses of differences between groups by questionnaires were undertaken.

Item Analyses

Item analyses of risk and nonrisk group differences are presented in this section. The performance of the nonrisk group was believed to exceed that of the risk group on individual questionnaire items. To address this question, scoring of items in developmental areas on the eight original questionnaires (4, 8, 12, 16, 20, 24, 30, and 36 month intervals) was compared and a mean score for each was generated. A total of 240 items were examined.

Table 21 presents for each skill area the number of items in which 1) the mean score of the risk group exceeded the nonrisk group, 2) the mean score of the nonrisk group exceeded the risk group, and 3) the mean scores of the two groups were equal. The number of risk and nonrisk subjects included in this analysis was 6,377 and 1,682, respectively. From a total of 231 items, there were 19%, or only 43 items, in which mean scores for the risk group exceeded the mean scores for the nonrisk group.

Table 22 shows the same comparison by questionnaire interval. The number of items in which the risk group mean scores exceeded, were less

Table 21. Number of items by area on which the mean score for the risk group was greater than, less than, or equal to the nonrisk group

Area	Greater	Less	Equal
Communication	10	34	0
Gross motor	6	40	0
Fine motor	7	37	1
Problem solving	7	39	2
Personal-social	13	35	0
Total	43	185	3

Risk $n = 6,377$.

Nonrisk $n = 1,682$.

Note: Nine items in these developmental areas were changed between revisions and, therefore, were not included.

Table 22. Number of items by questionnaire in which the mean score for the risk group was greater than, less than, or equal to the nonrisk group

Age interval (in months)	n	Greater	Less	Equal
4	1,501	12	18	0
8	1,405	4	24	1
12	1,185	6	24	0
16	1,057	6	23	0
20	930	3	24	0
24	898	8	19	1
30	566	3	24	1
36	517	1	29	0
Total	8,059	43	185	3

Risk $n = 6,377$.

Nonrisk $n = 1,682$.

Note: Nine items in these developmental areas were changed between revisions and, therefore, were not included.

than, or equal to the nonrisk group are displayed for each questionnaire age interval. As can be seen in Table 22, the 43 items on which the risk group's mean score exceeded the nonrisk group are distributed across intervals with the greatest frequency at 4 months.

Analysis of Group Differences for the Risk and Nonrisk Groups

To test for differences between the risk and nonrisk groups, the five area scores (e.g., communication, fine motor) for the questionnaire intervals (e.g., 4, 8, 12, 16, 20, and 24 months) were compared. A 2×6 (group \times questionnaire) mixed model multivariate analysis of variance (MANOVA) was conducted. Status of the subject as either risk or nonrisk was entered as a between-subjects factor, while questionnaire interval was entered as a within-subjects variable. Any cases with incomplete data for the within-subjects variable were deleted from this analysis, leaving 249 remaining cases spanning the 4 to 24 month intervals. Because cell sizes in this design were unequal, an unweighted means or regression approach was used in the analysis of these data (Tabachnick & Fidell, 1989) using SAS GLM (SAS Institute, 1990).

With the use of Wilks's criterion, the combined dependent variables were significantly affected only by the interaction of risk status and test interval, $F(25, 4567) = 2.27, p < .0005$, but not by either risk status or test interval independently. The results reflected a very small association between this interaction term and the combined dependent variables ($\eta^2 = .05$). Because the focus of this analysis was the relationship between risk status and scores on the questionnaires and the main effect of risk status was not significant, follow-up analyses were not performed.

These analyses suggested that differences between groups were minimal and not consistent across questionnaire age intervals or across areas. These analyses support the decision to combine risk and nonrisk groups for determining cutoff points as well as for determining the validity and reliability of the questionnaires.

Analyses of differences between nonrisk and risk groups yielded interesting comparisons. As predicted, the nonrisk group had more items whose mean score exceeded those for the risk group. However, results of the 2×6 (group \times questionnaire) MANOVA suggested no consistently significant differences between groups. The decision to combine risk and nonrisk groups to determine cutoff points using ROC analysis was further supported by multivariate analysis described previously.

REFERENCES

- Bayley, N. (1969). *Bayley Scales of Infant Development*. San Antonio, TX: The Psychological Corporation.
- Bricker, D., & Squires, J. (1989). The effectiveness of parental screening of at-risk infants: The infant monitoring questionnaires. *Topics in Early Childhood Special Education, 9*(3), 67-85.
- Brinker, R., Franzier, W., Lancelot, B., & Norman, J. (1989). Identifying infants from the inner city for early intervention. *Infants and Young Children, 2*(1), 49-58.
- Brown, W., & Brown, C. (1993). Defining eligibility for early intervention. In W. Brown, S.K. Thurman, & L.F. Pearl (Eds.), *Family-centered early intervention with infants and toddlers: Innovative cross-disciplinary approaches* (pp. 21-42). Baltimore: Paul H. Brookes Publishing Co.
- Cohen, M., & Gross, P. (1979). *Developmental resources: Behavioral sequences for assessment and program planning* (Vols. I & II). New York: Grune & Stratton.

- Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.
- Gradel, K., Thompson, M., & Sheehan, R. (1981). Parental and professional agreement in early childhood assessment. *Topics in Early Childhood Special Education*, 1, 31–40.
- Green, D., & Swets, J. (1966). *Signal detection theory and psychophysics*. New York: John Wiley & Sons.
- Hunt, J., & Paraskevopoulos, J. (1980). Children's psychological development as a function of the accuracy of their mothers' knowledge of their abilities. *Journal of Genetic Psychology*, 136, 285–298.
- Knobloch, H., Stevens, F., & Malone, A.F. (1980). *Manual of developmental diagnosis: The administration and interpretation of the Revised Gesell and Amatruda Developmental and Neurological Examination*. New York: Harper & Row.
- McCarthy, D. (1972). *McCarthy Scales of Children's Abilities*. San Antonio, TX: The Psychological Corporation.
- MECCA. (no date). *MinNesota Interactive Readability Approximation Program*. Lauderdale, MN: Author.
- Newborg, J., Stock, J.R., Wnek, L., Guidubaldi, J., & Svinicki, J. (1984). *Battelle Developmental Inventory*. Chicago: Riverside.
- Peterson, W., Birdsall, T., & Fox, W. (1954). Trans. IRE Prof. Group Inf. Theory PGIT-4, 171.
- SAS Institute, Inc. (1990). *SAS/STAT user's guide* (Vol. 2, Vers. 6). Cary, NC: Author.
- Tabachnick, R., & Fidell, L. (1989). *Using multivariate statistics* (2nd ed.). New York: Harper & Row.
- Thorndike, R., Hagen, E., & Sattler, J. (1985). *Stanford-Binet Intelligence Scale* (4th ed.). Chicago: Riverside.