matrix is not supplied but will be part of a new MSP version to be released in 1993. Molenaar and Sijtsma (1988) generalized the method for reliability estimation to the total score based on polytomous items. This method assumes that the ISRFs of different items do not intersect.

APPLICATION OF MOKKEN SCALE ANALYSIS TO TRANSITIVITY TASKS

Transitivity Tasks, Scoring Rules, Items, and Population

The purpose of this section is to illustrate how Mokken scale analysis can be used to construct a scale for the ability of transitive inference. The results from this application are compared with the results from a Rasch analysis of the same data.

Transitivity tasks. The concept of transitivity is well known from the theory by Piaget (Piaget & Inhelder, 1941; Piaget & Szeminska, 1941). In this study, transitivity tasks consist of three or four objects (e.g., sticks, tubes, balls, cubes, or disks) that should be compared with respect to one physical property (e.g., length, weight, or size). The tasks used here were derived from Piaget and Inhelder; Piaget and Szeminska; Piaget, Inhelder, and Szeminska (1948); Brainerd (1974); and Harris and Bassett (1975). In each transitivity task, the objects were presented pairwise. For example, in a task consisting of three objects, the subject was allowed to compare the objects within two of the three possible pairs directly without the use of measurement instruments. Next, all three objects were presented simultaneously and the subject was asked to infer the transitive relation between the objects from the third pair that were not compared directly. After the response was recorded, the subject was requested to justify it in case it was a correct response.

Scoring rules. Piaget and his associates (Piaget & Inhelder, 1941; Piaget & Szeminska, 1941) considered the evaluation of the justification given for a correct inference (judgment-plus scoring rule) to be adequate to determine whether a cognitive structure concerning transitive inference is present. In contrast, Brainerd (1973) and Kingma (1981, 1984) considered the evaluation of the inference (judgment-only scoring rule) more suitable for inferring the presence of a cognitive structure. In this study, the scalability of the transitivity tasks was investigated under both scoring rules using Mokken scale analysis for dichotomous items. Next, these scoring rules were combined into a new scoring rule that yields trichotomous item

scores. The resulting data were analyzed using Mokken scale analysis for polytomous items.

Because it is a controversial topic, it seems appropriate to give special attention to the rationale underlying different scoring rules for transitivity tasks. According to Piaget (1942; Piaget & Inhelder, 1941; Piaget & Szeminska, 1941), a cognitive structure is absent if a subject is unable to give a correct inference or if the inference is correct by coincidence. In the latter case, the subject did not make all comparisons that are necessary for the transitive inference but somehow arrived at the correct solution, for example, as a result of guessing or using an inadequate strategy. A cognitive structure is partially present if a subject gives a correct inference that is based on direct verification. For example, in spite of the instruction to leave the objects on the table, the subject picks up two objects and compares them to verify his or her statement about the transitive relation. A cognitive structure is also partially present if the inference is based on perceptual aspects of the objects. For example, the subject infers that C is longer than A because this property is visible to him, not because it is inferred logically from the comparison of the objects in the other pairs. A cognitive structure is operational if a subject gives a correct answer by producing a number of coherent transitive inferences. For example, A is shorter than B, and B is shorter than C; therefore, A is the shortest.

Based on these considerations, three scoring rules were used. The judgment-only rule yields an item score of zero for an incorrect inference and a score of unity if the inference was correct. The judgment-plus rule differs from the judgment-only rule in that a score of unity was obtained only if, in addition to a correct inference, a correct justification for it was given. Obviously, the judgment-plus rule reduces the impact of measurement errors and strategy errors on the data. The third scoring rule combines the information about inference and justification in a trichotomous item score. In fact, the data matrix obtained from the application of the combination scoring rule can also be obtained by summation of the data matrices obtained from the application of the first two scoring rules. A zero score thus reflects that the inference was incorrect. A partial credit score equal to unity reflects that the subject was unable to provide an explanation for his or her correct inference or that the subject based his or her correct inference on direct verification or perceptual aspects of a task. A score equal to 2 means that, in addition to a correct inference, a correct justification was given.

For the analysis of the dichotomous item scores, the program Mokken Scale was used. This program provides the P^0 matrix in addition to the P matrix but only handles dichotomous items. MSP was used to analyze the trichotomous item scores. It provides only the P matrix, not the P^0 matrix, however.

TABLE 1
H_g Coefficients and p Values of Seven Tasks Selected Under the Judgment-Only Rule, and H_g Coefficients of Five Nonscale Tasks With Respect to the Seven Tasks Selected, and Their p Values

Task	H_g	*p Value*
Selected tasks		
IX	.50	.30
X	.52	.52
VII	.51	.84
III	.53	.88
I	.46	.94
VIII	.55	.97
VI	.59	.97
Tasks not admitted to the scale		
XII	−.20	.48
XI	−.03	.64
IV	−.07	.78
V	.20	.80
II	−.01	.81

Items and population. The test consisted of 10 transitivity tasks and 2 pseudotransitivity tasks that were added for validation purposes. The pseudotransitivity tasks resemble transitivity tasks but do not require a transitive inference for their solution. A description of each task seems beyond the scope of this study (for more details, see Verweij, Koops, & Sijtsma, 1992; Verweij, Sijtsma, & Koops, 1992).

Subjects were 425 Dutch pupils from 10 primary schools (Grades 2 through 6). Across grades, the mean ages expressed in months were 92.6, 102.7, 111.9, 126.1, and 139.9. From each grade about the same number of pupils was sampled. Both sexes were equally represented in each grade.

Results of Mokken Scale Analysis[1]

Analysis of judgment-only data. Using the item selection procedure with a lower bound H = .3, Mokken Scale selected seven tasks with H = .52. Table 1 shows the p values and the H_g coefficients of these tasks and also of the five tasks that were not admitted to the scale.

The rejection of the pseudotransitivity tasks, XI and XII, is in agreement with the assumption that these tasks do not require a transitive inference for their solution. Tasks II and IV formed a separate scale (H = .35). In contrast with the other transitivity tasks, these tasks required the inference

[1]The raw data used in this study can be obtained from the authors for reanalysis purposes. Requests should be sent to the first author.

of an equality rather than an inequality. Based on their negative H_g values with respect to the seven scaled items, it would seem plausible to conclude that the inferences of an equality or an inequality require different abilities. Although positive, the scalability coefficient of Task V did not exceed the lower bound $H_g = .3$. Verweij, Koops, and Sijtsma (1992) argued that this low item value was caused by an order effect due to the particular presentation order of the objects within this item. Task V was the only task in which the objects were presented in an ascending order (with respect to weight). The objects of the other tasks were presented in a descending order or required an inference of an equality (Tasks II and IV).

No significant deviations from the expected ordering in the rows and columns in the P matrix and the P^0 matrix were found (Table 2). Thus, assuming that the DM model holds, the reliability of the total score based on seven tasks had an estimated reliability of .68.

Analysis of judgment-plus data. The 10 transitivity tasks were admitted to the same scale with H = .76. All item coefficients were larger than .5 and six of them had values higher than .7 (Table 3). The pseudo-transitivity tasks were rejected. This excellent result can be explained by noting that unity scores (correct solutions) in the judgment-only data that resulted from guessing or from following an inadequate strategy were replaced by zero scores, thus yielding the judgment-plus data matrix. For

TABLE 2
P and P^0 Matrices for Seven Tasks Selected Under the Judgment-Only Rule

	Task						
Task	*IX*	*X*	*VII*	*III*	*I*	*VIII*	*VI*
P Matrix							
IX	–	.22	.28	.28	.29	.30	.30
X	.22	–	.48	.49	.51	.51	.52
VII	.28	.48	–	.79	.82	.83	.84
III	.28	.49	.79	–	.86	.88	.88
I	.29	.51	.82	.86	–	.92	.93
VIII	.30	.51	.83	.88	.92	–	.96
VI	.30	.52	.84	.88	.93	.96	–
P^0 Matrix							
IX	–	.40	.13	.10	.05	.03	.02
X	.40	–	.12	.09	.04	.03	.02
VII	.13	.12	–	.06	.03	.02	.02
III	.10	.09	.06	–	.04	.02	.02
I	.05	.04	.03	.04	–	.01	.01
VIII	.03	.03	.02	.02	.01	–	.01
VI	.02	.02	.02	.02	.01	.01	–

TABLE 3
H_g Coefficients and p Values of 10 Tasks Selected Under the Judgment-Plus Rule, H_g Coefficients of 2 Nonscale Tasks With Respect to the 10 Tasks Selected, and Their p Values

Task	H_g	*p Value*
Selected tasks		
II	.54	.03
IV	.54	.04
IX	.52	.06
X	.69	.09
V	.79	.18
I	.79	.28
III	.82	.36
VI	.80	.53
VII	.78	.62
VIII	.78	.72
Tasks not admitted to the scale		
XI	.00	.50
XII	.03	.11

Tasks II and IV, forming a separate scale under the judgment-only scoring rule, this would imply that many subjects tended to use the same inadequate strategy based on direct verification or perceptual aspects on both tasks. The elimination of many measurement and strategy errors also much improved the scalability of Task V. It may be concluded that the justification scores provide more accurate information about the process leading to a solution than the mere correct–incorrect scores.

No significant deviations from the expected orderings were present in the P and P^0 matrices. Because of the limited value of showing these matrices for each analysis, they are not presented here. The total score based on 10 transitivity tasks had an estimated reliability equal to .89.

Analysis of trichotomous data. MSP selected eight tasks with H = .52. In Table 4, the H_g coefficients and the mean scores of these tasks and the four tasks that were not admitted to this scale are given. The pseudotransitivity tasks, XI and XII, were rejected from the scale. Exactly as with the judgment-only analysis, Task II and Task IV formed a separate scale with H = .38. This result may be explained by the apparent consistency of the use of inadequate strategies on both tasks as reflected by the partial credit score category.

In the P matrix of the eight tasks selected, significant deviations from the expected orderings in rows and columns were found for the first item step of Task IX. After removal of this task, it could be concluded that the seven remaining tasks allow the ordering of persons and items. The reliability of the total score, based on these seven tasks, was estimated to be .84.

TABLE 4
H_g Coefficients and Mean Scores of Eight Tasks Selected Under the Combination Score, H_g Coefficients of Four Nonscale Tasks With Respect to the Eight Tasks Selected, and Their Mean Scores

Task	H_g	*Mean Score*
Selected tasks		
IX	.36	.36
X	.41	.61
V	.46	.99
I	.61	1.22
III	.59	1.24
VII	.56	1.46
VI	.59	1.50
VIII	.57	1.69
Tasks not admitted to the scale		
XII	−.13	.58
IV	.16	.83
II	.09	.84
XI	.02	1.14

Results of Rasch Analysis

The dichotomous judgment-only and judgment-plus data sets were reanalyzed using the computer program PML (Gustafsson, 1977) for Rasch analysis. PML is suited for dichotomous item scores only. The Andersen (1973) chi-square test was used to test the null-hypothesis that all k IRFs are monotonic nondecreasing with the same slope. Molenaar's (1983) standard normal item statistic U_g was used to test for each individual task the null-hypothesis that the observed IRF equals the IRF expected, given the Rasch model. Finally, Van den Wollenberg's (1982) splitter item technique based on Andersen's test was used to test the null-hypothesis that a set of tasks is unidimensional. An extensive treatment of these methods and an application to empirical data is provided by Meijer et al. (1990). Only the 10 transitivity tasks were reanalyzed.

Analysis of judgment-only data. For the 10 transitivity tasks, the null-hypothesis of monotonic nondecreasing IRFs with equal slopes was rejected for a division of the sample into grades, but not for divisions into sex or age groups (Table 5). The U_g values shown at the bottom of Table 5 ($k = 10$) indicate for Tasks II and IV that their IRFs were flatter than expected (significantly positive U_g values at a 5% level). After removal of these 2 tasks, the newly determined U_g values for the remaining 8 tasks (not shown in Table 5) indicated that the IRF of Task V also was flatter than

TABLE 5
Rasch Analysis of Judgment-Only Data Using Andersen's Test for Item Sets and Molenaar's U_g Test for Individual Items

			χ^2	*df*	*p*
k = 10	Andersen	Grades 2 through 6 (5 groups)	71.5	36	.000
		Age (low-high)	6.8	9	.654
		Sex	8.1	9	.526
		Splitter item X	60.8	8	.000
k = 7	Andersen	Grades 2, 3 and 4, 5, 6 (2 groups)	6.0	6	.423
		Age (low-high)	7.5	6	.274
		Sex	8.5	6	.205
		Splitter item X	2.3	5	.805

Molenaar's U_g test	*Task*									
Item #	*VII*	*II*	*VI*	*IV*	*VIII*	*I*	*III*	*V*	*IX*	*X*
k = 10	−1.9	2.3	−1.5	4.7	−2.4	−.4	−2.1	1.3	1.0	−.5
k = 7	−.0		−.8		−.5	.4	−.2		2.2	.5

expected. Leaving out this third task and reanalyzing the remaining data yielded the Andersen test results for k = 7 in Table 5 and the U_g values found at the bottom line of this table. Given these results, it was concluded that the remaining 7 tasks had increasing IRFs with equal slopes.

For k = 10, the splitter item method yielded a rejection of the null-hypothesis of unidimensionality (Table 5). For k = 8 and k = 7 (the latter result is shown in Table 5), this hypothesis could not be rejected.

The combination of all test results leads to the conclusion that seven transitivity tasks conform to the Rasch model.

Analysis of judgment-plus data. From Table 6, it can be concluded that the null-hypothesis of monotonic nondecreasing IRFs with equal slopes cannot be rejected for any of the divisions of the sample. Because this result pertains to the complete set of tasks, the U_g values were not used to reject items.

The splitter item method led to a rejection of the null-hypothesis of unidimensionality for all four splitter items used. Following a graphical procedure described by Van den Wollenberg (1982) and Molenaar (1983), it was concluded that two subsets of tasks can be distinguished: Tasks I, III, V, VI, VII, and VIII all consist of three objects, and Tasks II, IV, IX, and X all consist of four objects. For splitter item VIII, Figure 1 shows that different item difficulties were estimated in two different subgroups, a result that violates the property of specifically objective item measurement. More specifically, the five items that are related to the splitter item, as

TABLE 6
Rasch Analysis of Judgment-Plus Data Using Andersen's Test for Item Sets and Molenaar's U_g Test for Individual Items

			χ^2	*df*	*p*
k = 10	Andersen	Grades 2, 3 and 4, 5, 6 (2 groups)	11.6	9	.235
		Age (low-high)	16.4	9	.060
		Sex	7.8	9	.554
		Splitter item III	35.5	8	.000
		VI	40.6	8	.000
		VII	32.6	8	.000
		VIII	39.8	8	.000

Molenaar's U_g test	*Task*									
Item #	*VII*	*II*	*VI*	*IV*	*VIII*	*I*	*III*	*V*	*IX*	*X*
k = 10	1.6	−.3	1.1	1.8	−.3	−2.8	−2.0	−.9	1.3	.3

concerns their number of objects, were more difficult in the subgroup with an incorrect response on the splitter item than in the subgroup with a correct response. This is the key characteristic of the splitter item method that leads to the conclusion that multidimensionality applies (for further details, see Van den Wollenberg, 1982).

DISCUSSION

The application of the Mokken models and the Rasch model to empirical data sets for transitivity tasks obtained under three different scoring rules led to several interesting results.

First, using three different scoring rules, Mokken scale analysis yielded strong scales for transitive inference in each case. For the judgment-only data, 7 tasks were selected that were in accordance with the MH model and the stronger DM model. Because the judgment-plus data contain fewer measurement and strategy errors, more tasks were expected to constitute the final scale. In fact, all 10 transitivity tasks were in accordance with both Mokken models. Because the data were more reliable and the number of items selected was larger, the reliability of the total score was higher for the judgment-plus scale than for the judgment-only scale. For the trichotomous data, 8 tasks were selected under the MH model, but under the DM model one of these tasks was removed from the scale. Although the number of tasks finally selected was smaller than for the judgment-plus data, the reliability was about the same, which is probably due to the larger number of score categories per task.

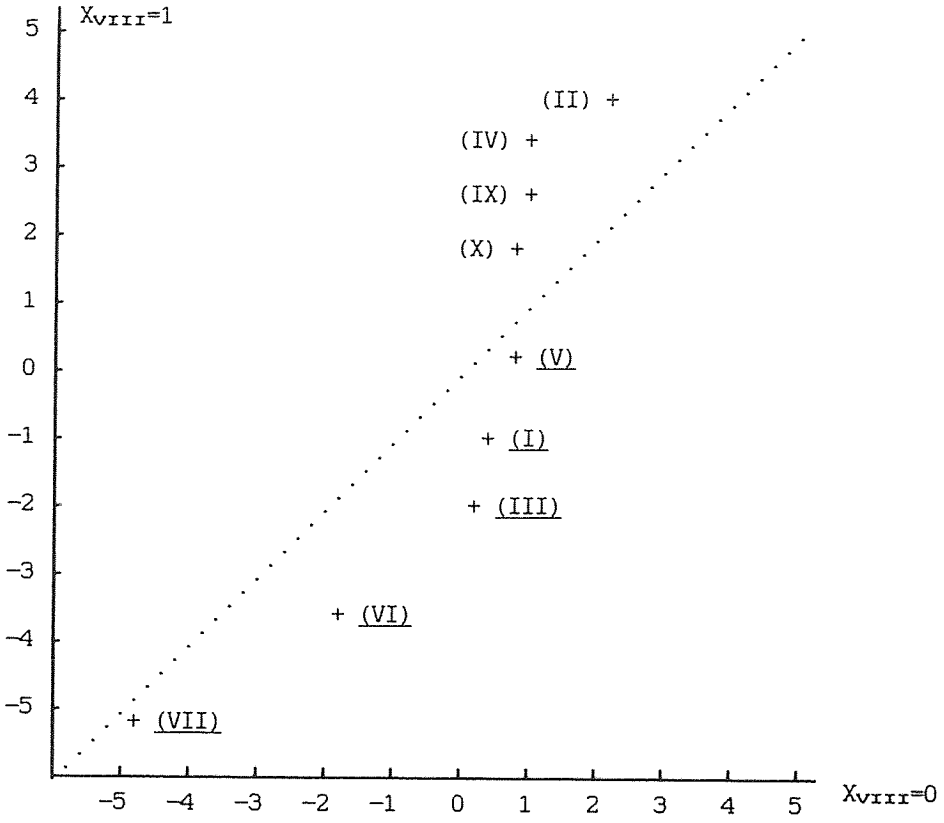

FIGURE 1 Splitter item method results for Task VIII. Item difficulty estimates in the subgroups having a zero or a unity score on Task VIII, respectively. Items consisting of three objects are underlined.

Second, given the controversial topic of which scoring rule to use with transitivity tasks, we suggested to combine the judgment-only and the judgment-plus rules into a new scoring rule. Not only does this new scoring rule combine information, it also corresponds to the distinction—between absence, partial presence, and presence of a cognitive structure—made by Piaget. Moreover, it was shown that with this combination scoring rule, it was possible to construct a useful transitivity scale.

Because the judgment-only rule quantifies correct responses due to measurement and strategy errors as unity scores, it seems to be the least preferable scoring rule. Furthermore, it seems hard to choose between the other two rules. Depending on the scoring rule used, either 7 (judgment-only), 10 (judgment-plus), or 8 (combination) tasks should be used for measurement of transitive inference according to the MH model. If, in

addition, an invariant item ordering is required, either 7, 10, or 7 tasks should be used, respectively. If the test is applied in another population rather than used in this study, new data should be collected and reanalyzed and other item selections may result. Assuming that the items predominantly measure transitive inference, finding different item selections in different populations would probably be due to changes in strategy or to bottom and ceiling effects for relatively young and old children, respectively.

Third, a reanalysis of the dichotomous data sets using the Rasch model led to the conclusion that, for the judgment-only data, the same 7 tasks were in accordance with this model as with the DM model. Because the DM model specifies IRFs to increase monotonically without intersections, several data sets satisfying this model will also satisfy the stronger Rasch model. This is not necessarily true, however, as was convincingly illustrated with the judgment-plus data analysis using the splitter item method: Although 10 transitivity tasks had logistic IRFs with equal slopes, application of this method revealed that a distinction must be made between processes underlying the solution of tasks with three or four objects, respectively. It is interesting to note ad hoc that the H_g coefficients (Table 3) also point in this direction: The 4 transitivity tasks with four objects have the lowest H_g values. Usually, variation in such relatively high H_g values is not taken as evidence of multidimensionality, however. It may be concluded that the fine-grained Rasch analysis, based on stronger assumptions, more explicitly leads to conclusions about multidimensionality than the weaker Mokken approach that only pursues ordinal measurement.

REFERENCES

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123–140.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick. *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Brainerd, C. J. (1973). Judgments and explanations as criteria for the presence of cognitive structures. *Psychological Bulletin, 79,* 172–179.

Brainerd, C. J. (1974). Training and transfer of transitivity, conservation, and class inclusion of length. *Child Development, 45,* 324–334.

Cliff, N. (1977). A theory of consistency of ordering generalizable to tailored testing. *Psychometrika, 42,* 375–401.

Croon, M. A. (1991). Investigating Mokken scalability of dichotomous items by means of ordinal latent class analysis. *British Journal of Mathematical and Statistical Psychology, 44,* 315–331.

Debets, P., Sijtsma, K., Brouwer, E., & Molenaar, I. W. (1989). MSP: A computer program for item analysis according to a nonparametric IRT approach. *Psychometrika, 54,* 534–536.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests* [Introduction to psychological test theory]. Bern: Huber.

Fischer, G. H. (1987). Applying the principles of specific objectivity and of generalizability to the measurement of change. *Psychometrika, 52,* 565–587.

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53,* 383–392.

Gustafsson, J. E. (1977). *The Rasch model for dichotomous items: Theory, applications and a computer program* (Internal Report No. 63). Sweden: University of Göteborg, Institute of Education.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 60–90). Princeton, NJ: Princeton University Press.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory.* Boston: Kluwer Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Beverly Hills, CA: Sage.

Harris, P. L., & Bassett, E. (1975). Transitive inference by four year old children. *Developmental Psychology, 11,* 875–876.

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika, 46,* 79–92.

Jong, A. de, & Molenaar, I. W. (1987). An application of Mokken's model for stochastic, cumulative scaling in psychiatric research. *Journal of Psychiatric Research, 21,* 137–149.

Kingma, J. (1981). *De ontwikkeling van quantitatieve en relationele begrippen bij kinderen van 4–12 jaar* [The development of quantitative and relational concepts in 4–12-year-old children]. Groningen, The Netherlands: Van Denderen.

Kingma, J. (1984). The sequence of development of transitivity, correspondence and seriation. *The Journal of Genetic Psychology, 144,* 271–284.

Kingma, J., & TenVergert, E. M. (1985). A nonparametric scale analysis of the development of conservation. *Applied Psychological Measurement, 9,* 375–387.

Lewis, C. (1983). Bayesian inference for latent abilities. In S. B. Anderson & J. S. Helmick (Eds.), *On educational testing* (pp. 224–251). San Francisco: Jossey-Bass.

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of 'scale analysis' and factor analysis. *Psychological Bulletin, 45,* 507–530.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement, 14,* 283–298.

Meredith, W. (1965). Some results based on a general stochastic model for mental tests. *Psychometrika, 30,* 419–440.

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* Berlin: De Gruyter.

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6,* 417–430.

Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to 'The Mokken scale: A critical discussion.' *Applied Psychological Measurement, 10,* 279–285.

Molenaar, I. W. (1982a). Een tweede weging van de Mokken schaal [A second weighing of the Mokken scaling procedure]. *Tijdschrift voor Onderwijsresearch, 7,* 172–181.

Molenaar, I. W. (1982b). Mokken scaling revisited. *Kwantitatieve Methoden, 3*(8), 145–164.

Molenaar, I. W. (1983). Some improved diagnostics for failure of the Rasch model. *Psychometrika, 48,* 49–72.

Molenaar, I. W. (1986). Een vingeroefening in item response theorie voor drie geordende antwoordcategorieën [An exercise in item response theory for three ordered response categories]. In G. F. Pikkemaat & J. J. A. Moors (Eds.), *Liber Amicorum Jaap Muilwijk* (pp. 39–57). Groningen, The Netherlands: Econometrisch Instituut.

Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden, 12*(37), 97–117.

Molenaar, I. W., & Sijtsma, K. (1988). Mokken's approach to reliability estimation extended to multicategory items. *Kwantitatieve Methoden, 9*(28), 115–126.

Niemöller, B. (1980). Mokken Scale. *STAP user's manual* (Vol. 4). Amsterdam: University of Amsterdam Press.

Piaget, J. (1942). *Classes, relations et nombres: Essai sur les groupements de la logistique et sur la réversibilité de la pensée.* Paris: Collin.

Piaget, J., & Inhelder, B. (1941). *Le développement des quantités chez l'enfant.* Neuchâtel: Delachaux et Niestlé.

Piaget, J., Inhelder, B., & Szeminska, A. (1948). *La géométrique spontanée de l'enfant.* Paris: Presses Universitaire de France.

Piaget, J., & Szeminska, A. (1941). *La genèse du nombre chez l'enfant.* Neuchâtel: Delachaux et Niestlé.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49,* 425–435.

Rosenbaum, P. R. (1987). Probability inequalities for latent scales. *British Journal of Mathematical and Statistical Psychology, 40,* 157–168.

Sijtsma, K., Debets, P., & Molenaar, I. W. (1990). Mokken scale analysis for polychotomous items: Theory, a computer program and an empirical application. *Quality & Quantity, 24,* 173–188.

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16,* 149–157.

Sijtsma, K., & Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika, 52,* 79–97.

Sijtsma, K., & Prins, P. M. (1986). Itemselectie in het Mokken model [Item selection in the Mokken model]. *Tijdschrift voor Onderwijsresearch, 11,* 121–129.

Van den Wollenberg, A. L. (1982). A simple and effective method to test the dimensionality axiom of the Rasch model. *Applied Psychological Measurement, 6,* 83–91.

Verweij, A. C., Koops, W., & Sijtsma, K. (1992). De constructie van een ontwikkelingspsychologische schaal voor transitiviteit [The construction of a developmental scale for transitivity]. *Nederlands Tijdschrift voor de Psychologie, 47,* 186–194.

Verweij, A. C., Sijtsma, K., & Koops, W. (1992). *A transitivity scale for longitudinal research.* Manuscript submitted for publication.

Wright, B. D., & Stone, M. H. (1979). *Best test design. Rasch measurement.* Chicago: MESA Press.