JOURNAL OF EDUCATIONAL MEASUREMENT
VOLUME 14, NUMBER 3 FALL 1977

MODELS, MEANINGS AND MISUNDERSTANDINGS: SOME ISSUES IN APPLYING RASCH'S THEORY

SUSAN E. WHITELY[1]
University of Kansas

Wright (1977) shows that a debate is developing between those who strongly advocate use of the Rasch model and those who have certain reservations about the extent to which the model meets some traditional concerns in trait measurement. In an earlier article, Whitely and Dawis (1974) presented the Rasch model in the context of least squares estimation, and noted some features that may limit the utility of the model in test development. Wright (1977) questions several of the specific interpretations and conclusions that were given in the earlier article. The current article is a response to those questions.

Although two areas of disagreement between Wright's (1977) and Whitely and Dawis' (1974) articles could be termed "pseudo-issues" (equivalent forms and technological sophistication), several areas represent real issues. The current article shows that 1) in practice, using least squares estimators as a first step in parameter estimations is neither awkward nor unnecessary; 2) correct interpretations and expressions for least squares standard errors were given in the Whitely and Dawis (1974) article, except for a minor typographical error; 3) large sample sizes are required for successful application of the model; and 4) some advantages of the model may be nullified in the process of meeting traditional goals in testing—namely validity and score interpretability.

Estimation

Several of Wright's (1977) criticisms are related to Whitely and Dawis' (1974) presentation of the Rasch model in the context of least squares estimation. Whitely and Dawis (1974) used the least squares approach (in contrast to the maximum likelihood method) for three reasons: 1) the sample-invariance of the parameters is more readily apparent in the least squares context; 2) the least squares procedure is easier to relate to the basic data matrix (presented by Whitely and Dawis, 1974, p. 165) so that the requirements of the data for the model may be highlighted; and 3) *JEM* readers may be more familiar with least squares estimation. In contrast, the disadvantage of the least squares presentation is that the reader may be led to favor least squares estimation, although maximum likelihood estimation should be more efficient.

Given this perspective, Wright (1977) and Whitely and Dawis (1974) appear to disagree on several estimation issues and these will be discussed.

Estimation procedures. Wright's (1977) conclusion what Whitely and Dawis' (1974) statements about the need for large sample sizes was prompted by their recommendation of an awkward and unnecessary two-step estimation procedure is inaccurate or misleading on several counts. First, as will be shown below, Whitely and Dawis' (1974) stress on large sample sizes stemmed from concerns other than estimation procedures.

[1]The author would like to thank Rene V. Dawis, Isaac J. Bejar, Kenneth O. Doyle and David L. Passmore for their comments on an earlier draft of this manuscript and David Thissen for his consultation on some of the technical issues. Although the viewpoints expressed in this manuscript are not necessarily those of these individuals, their assistance was invaluable.

Second, a two-step estimation procedure is neither awkward nor unnecessary. Maximum likelihood estimation is, in this application, an iterative procedure which requires the user to supply starting values for some parameters. Although any arbitrary set of values *can* be specified, some values, such as initial least squares estimates, are preferable in terms of computational efficiency and to avoid the possibility of terminating parameter estimation at a set of sub-optimal values (the problem of local minima or maxima). Third, in Wright and Panchapakesan (1969, pp. 36–37) maximum likelihood estimation is, in fact, presented as a two-step estimation procedure in which the least squares estimates for items "are then used as the initial values for the iterative procedure described in MAX," the maximum likelihood subroutine.

It should be noted, however, that Whitely and Dawis (1974) may be misread to mean that both initial item and person values are *required* to start maximum likelihood iterations. In fact, *either* set of estimates can be used efficiently to start the procedure.

Estimation formulas. A draft version of Wright (1977, pp. 222–223) contains a correction of Whitely and Dawis' rendering of the Wright and Panchapakesan formulas for the errors of the maximum likelihood parameter estimates. In fact, neither Wright (1976) nor Whitely and Dawis (1974) present formulas which could be correctly rendered from Wright and Panchapakesan's (1969) Formula 29. First, consistent with their least squares presentation of the Rasch model, the Whitely and Dawis' formulas for standard errors were correctly obtained from Wright and Panchapakesan's formulas 12 and 13[2] for *least squares error,* not maximum likelihood, and were mislabeled only due to the printer's inadvertent omission of the square root notation that was given on the galley proof. The Whitely and Dawis interpretations of the computational procedures for the Rasch parameter error terms correctly apply to the least squares formulas given in the article. Furthermore, although Wright's (1977) Formula 2 is correct for the standard error of the maximum likelihood item estimates, Formula 1 is incomplete. The standard error of maximum likelihood ability estimates (from Wright and Panchapakesan's Formula 29) contains a second term which increases the error for imprecision in item calibrations. Omission of this term is no small matter, for the formula is formally incorrect and the missing term can account for as much as 20% of the error variance.[3]

SAMPLE SIZE

Whitely and Dawis (1974, p. 169) stated the following conclusion about the sample size required to apply the Rasch model: "although the P_{ij}'s from the extremes can be estimated from the model, the need for very large N's in test development should be obvious." Although this statement shows that the possibility of some empty score levels

[2]Incidentally, the Wright and Panchapakesan (1969) Formula 13, the variance of ability estimates, is also mislabeled as a standard error.

[3]The complete formula for the standard error of the Wright and Panchapakesan (1969) maximum likelihood ability estimates is as follows:

$$Se(b_r) = \left[\frac{1}{\sum_i P_{ji}(1 - P_{ji})} + \frac{\sum_i (V(d_i)(P_{ji}(1 - P_{ji}))^2}{\sum_i (P_{ji}(1 - P_{ji}))^2} \right]^{1/2}$$

is *not* at issue, Wright's (1977, p. 219) statement that "the Rasch model can be and has been productively applied to sets of data as small as 100 persons" shows that sample size is a real issue. The key to understanding the difference between Wright's and Whitely and Dawis' account of sample size is the differing importance attached to a powerful test of fit of the data to the Rasch model, prior to having useful estimates of the parameters.

Importance of testing fit. A major advantage of a successful application of the Rasch model is being able to specify that the observed test response data arises from the interaction of a person's ability on the latent trait being measured, and the item's easiness; that is, the data fit the Rasch model. Additionally, the major features of the Rasch model—1) the independence of person measurement from the particular items used and 2) the independence of item calibrations from the particular persons sampled—formally depend upon the test data fitting the model. With respect to estimating parameters for existing tests, the several studies which apply a reasonably stringent test of fit are notable for the frequency with which the model is found to be inappropriate (i.e., Birnbaum, 1968; Brooks, 1965; Kearney, 1966). Thus, it is not reasonable to assume *a priori* that all or most test data fit the model.

Although a model may still be useful when its assumptions are not strictly met, few guidelines are available in the published literature about how various types and degrees of departures from the Rasch model's assumptions influence the usefulness of the parameter calibrations. Initial results on person calibrations (Whitely & Dawis, 1974; Wright, 1968) have indicated robustness for moderate departures from unidimensionality and equivalent slopes. Probably more serious, however, are departures from local independence of items [obtained from difficulty-ordering and position effects (i.e., Sax & Karr, 1962) and test context effects (Whitely & Dawis, 1976)], but it appears that no published research has examined this problem. For person-free item calibrations, little robustness would be *expected* for departures from unidimensionality and equivalent slopes when calibrating samples differ widely in ability. Again, unfortunately, no published results appear to address this issue directly, and neither Wright's (1968) data on person-free *test* calibration nor Anderson, Kearney and Everett's (1968) item comparisons are adequate substitutes. Wright's (1968) paper concerns the likelihood ratios associated with the various total scores, rather than the equivalence of item parameter estimates, while the ability distributions in Anderson *et al*'s (1968) two populations probably did not vary greatly, if at all.

Thus, at least some major advantages of the Rasch model depend, either directly or indirectly, on fit of the data to the model, and the test developer cannot wisely assume hir data meet the requirements without administering proper tests of fit.

Testing fit. Wright (1977) explores the problem of sample size by examining the effect of N on the log likelihood standard errors of item calibration differences, using a two-group comparison of the maximum likelihood item estimates. This method must be questioned for two reasons. First, the method is not really an adequate test of the requirements of the data. Two-group comparisons are more appropriate for checking the appropriateness of the model in different populations. Items which fit the Rasch model should have comparable likelihood estimates obtained from *each* score level. By using only two effective score levels in the two-group comparison, one may sum over some significant departures, particularly at the extremes of the distribution. Second, reporting standard errors on the scale of item easiness, E_i, rather than for logarithms,

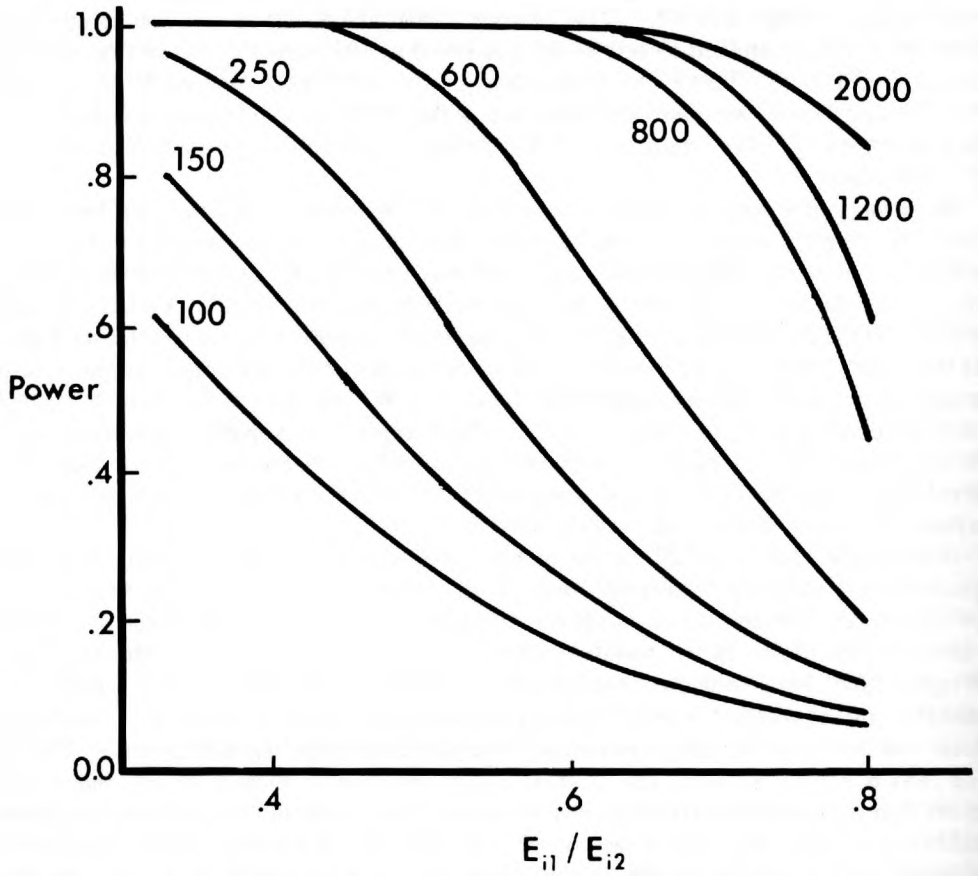

Figure 1. Power of the Two-Group Test for Item Calibration Differences for Several Sample Sizes.

d_i, would give a somewhat different impression of the size of the error (i.e., it would seem larger). Third, and most important, the two-group comparison has little power to detect sizable degrees of departure from the model for the sample sizes indicated by Wright.

Figure 1 presents the power of Wright's two group test (presumably a t-test) of item differences for various sample sizes, under maximum likelihood estimation. The abscissa represents some hypothetical true differences in item parameters between two groups. Each value represents the ratio, E_{i1}/E_{i2}, of the simple likelihood ratios for solving item i, where E_{i1} and E_{i2} are obtained separately from Group 1 and Group 2. The ratio E_{i1}/E_{i2} is the antilog of the difference between the log likelihoods ($\ln E_{i1} - \ln E_{i2}$) in Wright's two-group comparison. The standard errors for item differences (log scale) were computed from the formula used by Wright (1977) to examine sample size, as follows:

$$SED = \sqrt{2}\ \sqrt{6/(N/2)}$$

where N denotes examinee sample size for each group. It was assumed that item differ-

ences were distributed around the hypothetical true values as a non-central t, with the standard error given above. However, the power calculations were actually obtained from the standard normal distribution, as it is an adequate approximation to the non-central t for the sample sizes presented here. The power calculations are for a two-tail test at $\alpha = .05$.

From Figure 1, it can be seen that for an N of 100, the test has little power to detect differences even when the likelihood ratio of solving an item in Group 1 is only .33 that of Group 2. Even with an N of 250, the power to detect moderate item differences is low, e.g., power is about .28 when the item likelihood ratio from Group 1 is .66 of the likelihood ratio in Group 2. In fact, the power for a likelihood ratio of .66 does not reach the .90's until an N of about 800 is obtained.

Thus, it is clear that the two-group test of fit needs large N's if a moderate degree of departure of the data from the Rasch model is to be detected. Since this test is not really adequate for the reasons cited above, a better test, with more effective score groups, will probably require even larger N's.

Conclusion. Given the importance of testing fit, and the need for a reasonably powerful statistical test, successful application of the Rasch model requires large sample sizes at some phase in the test development process. Since the power of a test of fit is dependent on N, the choice of sample size should be guided by the degree of departure from the model that the test developer wishes to detect. At the extremes, a sample of several thousand can detect trivial departures, while a small N (less than 800) fails to detect sizeable differences.

OBJECTIVE MEASUREMENT REVISITED

Whitely and Dawis were concerned about the properties of objective measurement which were gained by using the Rasch model, and stressed that the Rasch model's objectivity would not be realized unless it was inherent in the test data. Wright (1977, p. 220) agrees with this, but also states that lack of fit to the model may imply that the data are not suited for any kind of measurement "if the measurement sought is to be objective, in the sense ordinarily meant when scientists 'measure'." This issue needs further elaboration for, carried to its extreme, the need for "objectivity" could imply not only that fit to the Rasch model should become a standard in test development, and that tests which do not fit the model are not "scientific." Since it was stated above that many otherwise reputable tests do not fit the model, this requirement would be no small matter.

Unfortunately, the term "objectivity" has many associated meanings in addition to that implied by successful calibration with the Rasch model. Use of the term in yet another way may perpetuate some misconceptions about the nature of the calibrated measure. This problem, in part, is what prompted Whitely and Dawis' (1974) warning of "superficial objectivity" in Rasch-calibrated data. Wright (1968, p. 87) clarifies his use of the term "objectivity," as achieved with the Rasch model, by defining two necessary conditions:

> First, the calibration of measuring instruments must be independent of those objects that happen to be used for calibration. Second, the measurement of objects must be independent of the instrument that happens to be used for measuring.

Wright's use of the term will be referred to as "specific objectivity." Two general ques-

tions must be asked to guide the test developer in determining the value of the specific objectivity offered by fitting the Rasch model: 1) Is specific objectivity essential to measuring psychological characteristics, so that other goals of measurement become secondary? and 2) Does the achievement of specific objectivity influence other goals in test construction (i.e., validity in particular)?

Specific objectivity and essential qualities of tests. Specific objectivity describes the kind of information that is represented in test scores. To evaluate how essential specific objectivity is, it is necessary to consider more generally the kind of information that may be represented by measurements. Torgerson (1958, pp. 21–37) presents one of the more extensive discussions available. Basing his discussion on the work of the philosopher N. R. Campbell, Torgerson defines three kinds of measurement: 1) *fundamental measurement,* in which numbers represent the calibration of the property according to natural (empirical) laws and do not necessitate the measurement of any other variable, 2) *derived measurement,* in which the numbers representing the property obtain meaning from a precise relationship to other properties, and 3) *measurement by fiat,* in which the presentation of the property depends on presumed relationships between the observations and the concept of interest. Torgerson dismisses derived measurement from consideration in the social sciences, since the laws and theories are not precisely formulated. Fundamental measurement, on the other hand, was seen as having applicability in some areas of psychology, particularly psychophysics, since the numerical values for the various quantities can be assigned by an experiment, such as a judgmental task. Torgerson classifies mental tests as measurement by fiat, since the numbers are not assigned by an experimental process, nor is the quality to be calibrated (level on the trait) directly measured.

Since mental tests must be classified as measurement by fiat, the quality of a test must be evaluated by appealing to criteria outside the measuring process. These criteria have been specifically elaborated in terms of validity, especially construct validity, for traits, and are considered an essential aspect of a mental test. If we apply Torgerson's distinctions about the kinds of information in a measure, specific objectivity can be seen as essential for a fundamentally-measured variable, since meaning here depends on the calibration process itself.

That the attainment of specific objectivity bears no necessary relationship to validity can be shown by extreme examples. The calibration of a test may come from a very arbitrary process of assigning weights to response patterns, and, although it may be inefficient to work this way, the scale will be valid to the extent that it enters into a number of important relationships with other variables, in accordance with a nomothetic network (Cronbach & Meehl, 1955). In contrast, the calibration of a measuring instrument may rigorously meet the criteria of specific objectivity, but bear no relationship to the trait purportedly measured. Therefore, although specific objectivity may be a desirable property, it is neither an essential component nor a substitute for validity.

The interaction of validity and specific objectivity. The influence of developing a test to fit the Rasch model on test validity has not yet been carefully studied. However, some expectations can be stated. In general, fitting the model can be expected to *narrow* the scope of a test. First, rigorously meeting the requirement of unidimensionality should imply the development of tests which provide less general inductive summaries of behavior. Traits such as introversion-extroversion and general intelligence

are measured from somewhat heterogeneous tests, but correlate with many aspects of behavior. Second, and more speculatively, selecting items with uniform item characteristic function slopes may alter what is measured by a test if unidimensionality does not strictly hold in the full item domain. Third, the requirement of local independence may redefine what qualities can or cannot be measured. For instance, tests for which there are practice effects or learning from prior items will not fit the requirements of the model. Since many traits are supposedly related to learning, it may not be reasonable to exclude locally *dependent* item domains. In fact, successful measurement of a trait may depend on having such items, and, if so, test models should be built to meet the demands of substantive theory.

Conclusion. The properties achieved by the Rasch model are essential for fundamentally-measured variables, as are found in psychophysics. However, these properties are not essential in mental test measurements, which must rely on relationships outside the measuring process to obtain meaning. Although the advantage of having tests which satisfy a precise model should not be discounted, developing items to fit the Rasch model may result in tests which are objective only in a narrow sense, and at the possible expense of essential classical standards for trait measurement. Simply stated, data on the internal structure of a test may not be substituted for other kinds of validity data. Furthermore, selecting items to fit a very restrictive model may lead to narrow tests with few significant relationships to other variables.

INTERPRETABILITY OF RASCH ABILITY PARAMETERS

Differing perspectives on specific objectivity are also involved in the discrepancy between Wright's and Whitely and Dawis' account of the importance of anchoring Rasch parameters.[4] Wright (1977, p. 221) views anchoring as "a trivial matter of establishing a reference point" in estimation of the parameters. Whitely and Dawis (1974, p. 169) view anchoring as the "key to the sample-invariant interpretability of ability scores." The difference between these accounts is that Wright equates *score interpretability* with *parameter estimation* in this context. If the interpretability of scores derives solely from specific objectivity in the calibration process, as in fundamental measurement, then anchoring is a trivial event in the estimation process. However, if score interpretability depends on having scores which test users can interpret for examinees by referring to qualities outside the measuring process, anchoring is not a trivial problem. Traditionally, a meaningful score anchor may be either norm-referenced or domain-referenced. For a norm-referenced interpretation, the Rasch ability scores must be anchored to a relevant population. For a domain-referenced interpretation, the ability parameters must be anchored to a set of items which are intrinsically important to the attribute being measured. Achieving either of these qualities requires explicit concern during test development. Paradoxically, even if the Rasch parameters are meaningfully anchored in test development, they are inferior to classical standard scores for norm-referenced interpretations (the per-

[4]Whitely and Dawis (1974, p. 169) inadvertently omitted the word "geometric" from their description of the item anchoring procedure, using simple likelihoods. The geometric mean of the simple likelihoods, when set equal to 1.0, is identical to the continued product of simple likelihoods described by Wright (1976).

centile equivalents are not inherent in the score) and they have dubious meaning in a domain referenced interpretation (which may not be appropriate to trait measurement). But having an ability metric which adequately reproduces item responses would seem to be an advantage in score interpretation. Test users would benefit from further clarification of this issue.

PSEUDO-ISSUES

Wright (1977) questions Whitely and Dawis' (1974) description of statistically-equivalent item subsets from Rasch calibrations as providing a more limited equivalence than do classical parallel forms. In the discussion, Whitely and Dawis defined "limited" as not necessarily meeting some qualities of parallel forms, such as high precision and equal error variances, which are important to a test user who wants comparable information from different test forms. Wright, however, sees calibrated item subsets as being less limited, since any subset gives scores which can be converted into measures on the latent trait. These viewpoints are not really contradictory. When item subsets are not carefully selected for precision, then classical parallel forms will provide more equivalent information. However, when item subsets are carefully balanced to provide equivalent information for given score levels (equal information functions), then the information provided by the subsets is as comparable as that achieved from parallel forms, with the additional advantage that precision may be concentrated at specified score levels.

Another issue concerns technological sophistication. Wright (1977, p. 224) notes that computer-administered tests are not required to utilize the Rasch model, since paper and pencil tests "can be used for estimating measurements by means of simple tables of score-measure equivalents, without recourse to computers." Calibrating the Rasch-score equivalents or raw scores *is not* at issue on page 177 of the Whitely and Dawis (1974) article, for it was previously (pp. 164–165) stated that scoring tables are used after parameter calibrations have been obtained. Although the model has some interesting applications even for fixed content tests, the possibility of individualized testing, so that the "desired degree of precision for any person can be obtained from the fewest possible items" (Whitely & Dawis, 1974, p. 177) was seen as a major advantage of the model which is best implemented by computer. Testing may be individualized to a degree through a sequence of paper and pencil tests. However the immediacy and number of ability calibrations available greatly favor the interactive computer strategy.

GENERAL CONCLUSION

Whitely and Dawis (1974) were not particularly enthusiastic about the potential of the Rasch model to revolutionize test development. The reconsideration of the issues given here reaffirms that conclusion. Classical testing procedures have served test development admirably for several decades, and if a new model is to have impact it should offer alternatives to contemporary issues in applied testing, while still providing the major advantages of the classical model. Although the Rasch model has real potential for test efficiency, especially with individualized testing through computers, overemphasis of the model may inadvertently result in failure to achieve some important features of classical trait measurement.

REFERENCES

ANDERSON, J., KEARNEY, G.E., & EVERETT, A.V. An evaluation of Rasch's structural model for test items. *The British Journal of Mathematical and Statistical Psychology,* 1968, **21,** 231–238.

BIRNBAUM, A. Some latent trait models. In F. Lord & M. Novick (Eds.), *Statistical theories of mental test scores.* Reading, Mass.: Addison, Wesley & Co., 1968.

BROOKS, R. D. *An empirical investigation of the Rasch ratio-scale model for item difficulty indexes.* (Doctoral dissertation, University of Iowa, 1965). Ann Arbor, Michigan: University Microfilms No. 65-434.

CRONBACH, L. J., & MEEHL, P. E. Construct validity in psychological tests. *Psychological Bulletin,* 1955, **52,** 281–302.

KEARNEY, G. E. *Some aspects of the general cognitive ability of various aboriginal Australians as assessed by the Queensland Test.* Unpublished doctoral dissertation, University of Queensland, Australia, 1966.

SAX, G., & KARR, A. An investigation of response sets on altered parallel forms. *Educational and Psychological Measurements,* 1962, **22,** 371–376.

TORGERSON, W. S. *Theory and methods of scaling.* New York: John Wiley & Sons, 1958.

WHITELY, S. E., & DAWIS, R. V. The nature of the objectivity with the Rasch model. *Journal of Educational Measurement,* 1974, **11,** 163–178.

WHITELY, S. E., & DAWIS, R. V. The influence of test context on item difficulty. *Educational and Psychological Measurement,* 1976, **36,** 329–337.

WRIGHT, B. Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems.* Princeton, N.J.: Educational Testing Service, 1968, 85–101.

WRIGHT, B. Misunderstanding the Rasch model. *Journal of Educational Measurement,* 1977, **14,** (pp. 000–000).

WRIGHT, B., & PANCHAPAKESAN, N. A procedure for sample free item analysis. *Educational and Psychological Measurement,* 1969, **29,** 23–48.

AUTHOR

WHITELY, SUSAN E. *Address:* Department of Psychology, University of Kansas, Lawrence, KS 66045. *Title:* Assistant Professor of Psychology. *Degrees:* B.A., Ph.D. University of Minnesota. *Specialization:* Measurement; Individual Differences.