OFFICE OF SCALE RESEARCH

Technical Report
#0001

A Review of Selected Scale Construction and Evaluation Studies in Interpersonal and Organizational Trust

by

Grace J. Johnson

About the Author

Grace J. Johnson is a Doctoral student in the Department of Marketing at Southern Illinois University at Carbondale. Her academic background includes an undergraduate degree in Psychology and a Master's in Business. She worked in the Pharmaceutical industry and in International Business for several years before beginning to pursue her Ph.D. in Business. Her academic and work backgrounds have paved the way for her research interests, which include consumer behavior, the role of trust in e-Business, interactive marketing communication, and e-Medicine.

A Review of Selected Scale Construction and Evaluation Studies in Interpersonal and Organizational Trust

Trust, both interpersonal and organizational, has received a substantial amount of attention as a research topic, in recent years. To enhance the research in this area, a number of authors have come up with scales to measure trust. Different strategies can be used to develop personality scales. This paper examines twenty published studies on the topic of *scale construction and evaluation*, using as a focus the topic of *trust*. For each paper reviewed, a critical evaluation of the scale construction strategy used is offered, together with an assessment of the contribution of the study to the development of the trust construct. Following the 20 reviews, an overview section is presented, with general comments covering the cumulative work in the area of scales measuring trust.

A Review of Selected Scale Construction and Evaluation Studies in Interpersonal and Organizational Trust

The main objective of this paper is to examine a wide variety of published studies on the topic of *scale construction and evaluation*, using as a focus the topic of *trust*. Trust, both interpersonal and organizational, has received a substantial amount of attention as a research topic, in recent years. To enhance the research in this area, a number of authors have come up with scales to measure trust, and numerous published studies have focused on scale construction and evaluation. Different strategies can be used to develop personality scales. These are the *rational* strategy, the *empirical group discrimination* strategy, the *theoretical* strategy, and a *factor analytic* strategy. Once scales have been formulated, they have to be evaluated, during which various psychometric issues relating to different instruments are considered.

We have selected 17 articles in the area of scale construction, and 3 articles relating to further validation of scales already constructed. A total of 16 articles relate to trust itself directly. The remaining 4 relate to communal orientation, relational ethics, philosophies of human nature, and romantic love. These were included primarily because they had strong links to either the topic of trust, possessed subscales measuring trust, or because many trust studies had made use of items or subscales from them in their own scale construction. Many interpersonal trust scales measured this construct in the context of close romantic relationships, and hence the link to the one study measuring romantic love.

The reviews have been organized chronologically, with the further validation studies appearing toward the end, separately. (Refer to summary table, Appendix A). Each paper begins with a brief introduction to the main aim of the study, followed by an abbreviated description of the methods used. We have attempted to dwell in some length on a critical evaluation of each paper, balancing both strengths and weaknesses of the paper. Following the 20 reviews, we have presented a general overview section, with general comments covering all studies, and recommendations for further research in this area.

Wrightsman (1964)

The purpose of this paper was to develop a measure of *philosophies of human nature,* an age-old concept used in everyday life, but very new in the area of research. No attempt had been made to comprehensively measure this construct, although measures were available for specific aspects, such as the Machiavellianism Scale.

Since the concept was relatively new academically, very little was available by way of theoretical background. An analysis of historical and contemporary writings of theologians, philosophers, social scientists, and the mass media generated 6 basic dimensions constituting philosophies of human nature – trustworthiness, altruism, independence, strength of will, complexity, and variability. A few hypotheses were generated relating to how the 6 dimensions were related to each other.

For each hypothesized dimension, 20 statements were composed by the author from the previously cited sources. Half the items in each subscale were negatively worded. As a preliminary analysis of the scale, the 120 items were administered to 177 undergraduate students in 3 colleges. An item analysis was conducted based on its ability to discriminate between the top 25% and the bottom 25%, and 96 items were retained. This form was then administered another group of 100 graduate and 100 undergraduate students. Reliability - split-half and test-retest, with a gap of 3 months, was calculated. The next stage involved testing of hypotheses by administering the scale to 530 undergraduates. This group was also administered three other scales such as the Faith-in-People Scale and the Machiavellianism Scale. Reliability analysis revealed moderate to good values. Hypothesis testing results revealed general support for the hypotheses; for example, negative correlation between the scale and Machiavellianism, and this was taken to imply good discriminant validity. Distinct sex, age, and college differences were found, as the analyses were done separately. The researchers concluded that the scale appeared to have the reliability, validity, and differentiation required for research purposes.

The correct rational approach was taken in this study, as there was very little theory to support formulation of items. However, since only 20 items were generated for each dimension, there is no guarantee that the entire domain of items was covered. Considering the number of subjects available to them, the researchers could have conducted a detailed validation study, but failed to do so. A strong point was the separate analyses conducted for different colleges, sexes, and seniority (graduate/ undergraduate). This enabled them to detect differences that were there. Hypotheses were generated and tested, test-retest and internal consistency reliability analyses were conducted, and acquiescence was controlled for by negatively wording half the items. All these were strengths. The two major weaknesses of this study were the failure to conduct validation, though preliminary discriminant validation was done, and the failure to generate an item pool, and conduct item level analyses. Also, since the 6 dimensions were only hypothesized dimensions, an exploratory factor analysis could have been conducted to see if the dimensions fell out this way. The scale would be strengthened by further research in the directions indicated above.

Rotter (1967)

No motivation or objective is presented in this paper, but the objective appears very straightforward – development and initial validation of a scale to measure interpersonal trust (the Rotter Interpersonal Trust Scale). The trust construct used here is a *generalized expectancy* that the oral or written statements of other people can be relied upon.

A theoretical approach was used in the scale construction in this paper. Initially, interpersonal trust was defined, and its practical relevance explained in detail. Next, the construct was placed in an appropriate theoretical context by detailing the origin and development of the construct in social learning theory. The first step in scale construction involved generating a number of items and writing them up using a Likert-type format. Items were interspersed with filler items designed to disguise the purpose of the test. The experimental form of the test, together with the Marlowe-Crowne Social Desirability Scale, was then administered to a large group of introductory psychology students. Item level analysis was subsequently carried out, and an item was included if it had a good item-scale correlation, if it had a low correlation with the Social Desirability Scale, and if there was a good dispersion of responses across the 5 categories. Half the items were reverse scored. The final form of the test included 25 items measuring trust and 15 filler items. Both internal consistency and test-retest reliability were checked for and found acceptable. The next step involved testing the validity of the scale against two external criteria - observations of everyday behavior by using a set of sociometric scales, and self-ratings of trust. This involved administration of the test on a second group of students, together with the sociometric scales, and the Marlowe-Crowne scale. This analysis revealed good construct and discriminant validity for the Rotter Scale.

This is an example of a very good scale construction and evaluation study. Initially, some theoretical background was provided together with the author's definition of the construct. The theoretical background provided would have been strengthened had the author also provided a nomological net, relating the construct to other constructs. No information has been provided about how the initial item pool was generated, and how many items were present. This information would have indicated to the reader the adequacy and representativeness of the initial item pool. It was good that the researchers attempted to control for response style variance by using the Marlowe Crowne Scale (though some researchers have pointed out that this scale measures the need for approval rather than social desirability). Another strength is the item-level analysis that was carried out to determine content saturation, and the comparisons with the Social Desirability Scale. An index such as the Differential Reliability Index could have been used in this context, rather than mere comparisons of correlations, in order to reduce the variance in the item associated with content by that associated with desirability. Testing for reliability (both internal consistency and test-retest), was carried out, and this was procedurally very good, particularly as the researchers tracked the subjects over a long period of time in order to obtain test-retest reliabilities.

Finally, the preliminary construct and discriminant validation studies were good, and another source of strength. What could have been done subsequently was to carry out further validation studies to establish convergent validity also, for example with the use of a multitrait multimethod matrix. This could have been combined with initial testing of hypotheses to make this good study even better. Overall, this study was an excellent one, considering that it was written before papers such as Jackson's (1970) sequential system for personality scale development, which laid out step-by-step procedures for theoretical scale construction.

Rubin (1970)

The research described in this paper reports an attempt to improve the lack of available research into romantic love in the field of social psychology by introducing and validating a scale representing a social-psychological conception of romantic love. The researchers stated their intent to follow the strategy of construct validation recommended by Cronbach and Meehl (1955) in the whole process of defining romantic love, measuring it, and assessing its relationship to other variables.

Their conceptualization of this construct was a multifaceted one, as opposed to other theorists who viewed love as an emotion, need, etc. In order to assess the discriminant validity of the scale, it was constructed in conjunction with a parallel scale of liking.

The first step was the development of a large pool of items, based on popular notions in lay literature, and also on theory. Face validity of the items was assessed using two separate panels of student and faculty judges. 70 items were retained and administered to a group of undergraduates for pretesting. Separate factor analyses were run for men and women. In each case, one large, general factor was obtained, on which "love" scale items loaded, and a second, less important factor, on which "liking" scale items loaded. A subjective examination of the items at this stage revealed 3 separate dimensions of romantic love. In the next stage, the 26-item love and liking scales (no mention was made of what happened to the other items) were administered to 158 undergraduate dating couples. Item-scale correlations were calculated for each item with its own scale and with the other scale. The genders were analyzed separately. Internal consistency reliability was good, and good discrimination was found between the two scales. Low correlations were obtained with the Marlowe-Crowne Social Desirability Scale. Then, some predictive validation was done by testing some hypotheses in experimental settings, and these were found to be largely supported.

The approach followed here was a rational one, appropriate in this situation as there was lack of theory. Although the authors stated their intent to follow the process of construct validation, there was only limited evidence of this. Some discriminant validation was built in, right from the beginning, in designing liking and loving scales. And predictive validation was done toward the end. However, apart from this, there was little external validation with other scales or methods. Some strengths were the separate analyses for men and women, which showed up differences between the sexes, control for social desirability, and item-level analyses for content saturation. However, there were some glaring weaknesses also. The factor analyses produced one large general factor for the love scale items. Rotation of these factors might have shown up the three dimensions the researchers pointed out from a subjective evaluation of the factors. Dropping items from the 70-item measure was not explained or justified. Also, items with inappropriate patterns of correlations (higher correlation with other scale than its own scale) were not dropped. Finally, more work on convergent validation and test-retest reliability needs to be done to improve the scale.

Cook and Wall (1980)

The purpose of this paper is to develop three scales, for measuring the organizational variables, *trust*, *commitment*, and *fulfillment of personal needs*. At the time this paper was written, there were very few measures that dealt directly with these variables in an organizational setting.

Definitions for each variable were provided from the theoretical orientation taken by the authors, but there was no detailed exploration of theory or explanation of constructs and their relation with each other. The items for the scales were generated through two interview studies with blue-collar workers, all male, from a wide variety of industries in England, Scotland, and Wales. After the interviews, the authors generated the items guided by the interviews and the conceptual orientation taken by them. The questionnaires were then administered to volunteers, in two studies. After the first study, item-scale correlations were examined, and those items with very low correlations were dropped. After both studies were conducted, internal homogeneity was examined by means of coefficient alphas, and item-scale correlations. Also, t-tests were done to compare values on Studies 1 and 2. Test-retest reliability was calculated with a 6-month interval. All values were found to be acceptable.

Then, a factor analysis was conducted to see if the 3 factors were conceptually independent. Four factors were extracted, with one each for commitment, fulfillment of personal needs, and two for trust, which the authors distinguished as trust between peers, and management. In the second study, three other unrelated organizational measures had also been administered. Examination of the correlation matrices between all scales revealed good convergent and discriminant validity for the scales. The authors concluded that the measures were reliable, stable, and valid, but emphasized that further construct validation needed to be done.

At the time this paper was written, substantial theoretical groundwork had already been laid with respect to the three variables dealt with here. Because of this, there was no detailed theoretical overview preceding the development of the scales, although clear definitions were provided for each construct. Item generation was done on a theoretical-rational basis, as no mention was made of a large number of items being generated, for the item pool. Test-retest reliability, item content saturation analysis, discriminant validation, and internal consistency analyses were conducted. All these were excellent. However, some of the drawbacks pointed out here may help improve future studies conducted in this area. Firstly, all male subjects were used. Thus, the scales were valid only for use with male subjects, even though the researchers seemed to imply that their scales could be suitable for general use. A strength was their examination of discriminant and convergent validity with the use of the correlation matrix, but a more systematic way of doing this, such as with the use of a multitrait-multimethod matrix. Finally, the use of factor analysis was inappropriate; since they knew that there were three factors underlying, a confirmatory rather than exploratory method should have been used. Also, if a strong, theoretical approach, such as Jackson’s sequential approach had been used, there would have been no need to conduct factor analysis at all. Response-style variance should also have been controlled for, and the study would have been strengthened by separate analyses for distinct groups within their samples.

Larzelere and Huston (1980)

This study's aim was the operationalization of the concept of dyadic trust, demonstrating its measurement, and reporting on the relationship between trust and other related constructs. The need for this work was felt because existing measures of trust only measured *generalized trust* rather than trust in close human relationships (*dyadic trust*).

A rational strategy was used. The first part of the paper was devoted to a review of existing literature on trust, its relationship with other closely-related constructs, and defining the construct. The scale was generated by a rational approach of selecting 57 suitable items from a number of existing measures of trust, and modifying them for the specific focus in this study. The items were then administered to 120 females and 75 males who were involved in a close relationship such as dating couples, newly married or long-term married individuals, together with the Marlowe-Crowne Social Desirability Scale, two different measures of generalized trust, and measures of love, and extent of self-disclosure. Then, exploratory factor analysis was used to assess the unidimensionality of the item pool. It was concluded that dyadic trust was a unidimensional construct. Items were selected on the basis of 4 criteria – Jackson's DRI, distribution of responses across categories, repetitiveness of content, and direction of wording. Half the items were reverse scored. Item-total correlations were calculated and found to be good. Coefficient alphas were high, and correlations with Social Desirability were very low. Correlations with both generalized trust scales were found to be very low, thereby establishing good discriminant validity, according to the authors. Also, good discriminant validity was established with the love and self-disclosure scales. The authors concluded that the scales demonstrated adequate face validity, high reliability and excellent construct validity.

Procedurally, the construction of the scale as discussed here was very good. An attempt was made to define the construct and place it in an appropriate theoretical context.

Having done that, however, the authors went on to borrowing items from other trust scales and modifying them. Rather than doing this, they should have generated an item pool based on their definition and the theory they subscribed to. It is quite likely that their item pool of 57 items was not comprehensive enough to represent the entire domain. Response style variance was controlled for, and this was a strength, as also was their selection of items based on the DRI and content saturation. Internal consistency reliability was calculated, but not test-retest, and this needs to be done. Use of the rational strategy does result in items with good face validity, and good discriminant validity was demonstrated, but much more needs to be done in terms of convergent validity. A systematic way of attacking both problems simultaneously would be to conduct a multitrait multimethod analysis. The results would have been strengthened by separate analyses being conducted for women and men – their perceptions and responses to trust in close relationships might be very different. Overall, the study would have been a very good one had the researchers persevered with a theoretical strategy, rather than starting with a theory-based approach, and then moving onto a rational approach in generating items.

Johnson-George and Swap (1982)

The main objective of this paper was the construction and validation of a scale for the measurement of the varieties of interpersonal trust (the Specific Interpersonal Trust Scale – SITS) held by one individual for a specific other person. The need for such a scale was felt because:

a) Prior scales developed focused on measurement of *generalized* predispositions to trust, rather than trust in a specific other person or a specific type of trust. Such scales had been demonstrated to have limited usefulness in predicting trust except in highly ambiguous, novel, or unstructured situations, and did not accurately determine an individual’s trust in another under particular circumstances.

b) It was necessary to demonstrate differences between the trust construct and others closely related such as love, and liking.

A rational strategy was used in the construction of this scale. Initially, an item pool consisting of 50 items was generated by discussions with others and reviews of the theoretical literature dealing with interpersonal trust. The items dealt with hypothetical situations thought to involve trust in a specific other. Half the items were reverse keyed. They were presented to 15 judges, and they were asked to rate each item for its importance as a determinant of trust. 43 items with the highest interjudge agreement were retained. These items were randomly interspersed with Rubin's 13-item Liking and Loving scales (to assess discriminant validity), and administered to male and female undergraduate psychology students. Respondents were asked to think of a specific other person whom they trusted and answer the questions with respect to that person. Responses were subject to two levels principal components analysis – the first, of the full 69 items, in order to determine if trust, love and liking were discriminable constructs, and the second, with the 43 trust items. Separate analyses were conducted for males and females.

The results of the first factor analysis indicated that the loving, liking, and trust items formed separate factors. The authors therefore concluded that this was clear evidence for discriminant validity. The second factor analysis revealed different results for males and females For males, four interpretable factors were obtained and these were labeled a General Trust factor, Emotional trust, Reliableness, and Dependability. For females, three factors emerged: factor 1 seemed to be a blend of the male factors 3 and 4, and was labeled Reliableness. Factor 2 closely resembled the male factor 2 and was labeled Emotional Trust. Factor 3 was different from any of the male factors and was labeled Physical Trust. The authors felt that this factor was different because it reflected societal norms allowing females, more than males, to acknowledge their physical dependence on others. On this basis, two separate SITS scales, one for males (the SITS-M), and one for females (the SITS-F) were formed. The Male scale included the items that loaded on the first three factors, and were used as the Overall Trust subscale, the Emotional Trust subscale, and the Reliableness subscale. The Female scale included two subscales, one derived from Factor 1, as the Reliableness subscale, and the second derived from factor 2, labeled as the Emotional Trust subscale. There was some item overlap between the subscales. Each of the subscales demonstrated adequate reliability, ranging from .71 to .83. Some preliminary validation studies were then conducted, and these helped to establish discriminant validity.

The method used here in scale construction was a rational one – that is, the authors felt that the items chosen were rationally related to what they were trying to measure. This strategy is normally used when there is a lack of well laid out theory in a particular area. This was not the case with the trust construct. The history of research in this area extends to about 30 years prior to the writing of this paper. This paper could have benefited by the use of a theoretical, rather than a rational strategy. Particularly, the study would have benefited had some theoretical work been done in differentiating the trust construct from others that are conceptually similar, such as love, or liking, particularly as the authors pointed out that such work was missing in the literature. Also, some workable definition of trust, as they conceptualized it could have been provided.

Secondly, even with the use of the rational strategy, there were some weaknesses in their scale construction. Firstly, the item pool that they initially generated was very small – there were only 50 items, from which they dropped another 7. It is possible that such a small item pool might not be fully representative of the entire domain of items that could represent the trust construct. No attempts were made to control for response style variance. Also, although attempts were made to establish reliability (internal consistency), and discriminant validity, there is no real evidence that there was construct validity, convergent validity, item-level content saturation, test-retest reliability, or even, generally that there was any evidence that the scale was measuring what it was supposed to measure – interpersonal trust. The presence of item overlap could have inflated correlations.

Putting this aside, there are a few strengths in the paper, particularly in the way the study was conducted. For example, separate analyses were conducted for males and females that enabled them to isolate the differences between the sexes. Also, they made an attempt to establish discriminant validity right from the beginning by including the

Rubin’s Liking and Loving scale, and demonstrating that the items loaded on separate factors. Overall, if further work were done in the areas highlighted above, there could be marked improvements in the usefulness of the scale.

Clark, Ouellette, Powell, and Milberg (1987)

The authors of this paper were interested in formulating a new scale to measure Communal Orientation, as this was a relatively new construct in social psychology, and the authors wished to carry out some hypothesis testing in this area on the basis of individuals’ communal orientation. No such scales were available for use, and hence, the first part of the study involved construction of the scale and evaluating it.

Communal orientation was defined and a detailed review of existing empirical research in this area was conducted. However, very little by way of theoretical support was offered for this or any related constructs. On the basis of past research, a number of hypotheses were formulated, relating communal orientation to other constructs, and the authors set out to test these with the help of a scale which they constructed.

A rational strategy was used in scale development. Items on the Communal Orientation Scale (no details were given on how many items were generated) were pre-tested on 39 undergraduates, and on the basis of this pre-test, 14 descriptive statements were chosen, to which respondents had to reply on a 5-point scale. Half the items were reverse scored. The 14-item scale was then administered to a group of 561 college students. Cronbach’s alpha was found to be acceptable. The test was administered again, on another sample of 128 students for the purpose of calculating test-retest reliability, with an interval of 11 weeks, and this was found to be good. Item-scale correlations were found to range from .23 to .50 for the 14 items – the researchers accepted this value and claimed that items were not redundant. The test was administered once more on another sample of 565 undergraduates (no breakup given), and a principal components analysis was conducted with no objective stated for conducting it.

One large, general factor emerged, and two others were also selected based on the scree test and eigenvalue greater than 1 criterion. All 14 items loaded positively on factor 1, with most items' loading around 0.50. The three factors were labeled, with only the first one being considered relevant. Correlations between the Communal scale and others, such as the Marlowe Crowne Social Desirability Scale were compared. Low, insignificant correlation was observed with the Marlowe Crowne, and moderate, significant correlations observed with other measures of conceptually overlapping constructs such as social responsibility and emotional empathy. This was taken as implying good convergent validity, and the scale was used in subsequent research.

Frequently, when well-developed theory is not available about a particular construct or area, the rational strategy is used. This is the approach used in this study, as the area of research was relatively new. The items were generated on the basis of what the researchers thought were relevant, but this is no guarantee that the item pool was representative of the entire domain. Although items were selected for the final scale on the basis of a pre-test, the authors did not describe what criteria were used for item inclusion.

Some of the study's strengths related to correct following of procedures in scale construction, such as use of the social desirability scale to control for response style variance, reliability analysis (both internal consistency and test-retest), and the large number of subjects available to them. However, the flaws in their scale construction procedure far outweigh the strengths, and lead to questions about the validity of their scale. For example, the content saturation of the items was low, as shown by the item-scale correlations, even though they felt this was acceptable. The reason for the factor analysis is not at all clear, and it was also not done properly. When they obtained one large factor with all items loading moderately on it, they should have tried rotating the factors to get simple structure. Labeling the three factors and then dropping two of them was not correct and unacceptable, procedurally. Given the large number of student subjects that were available to them, they could easily have done a multitrait-multimethod kind of analysis, and this would have taken care of both convergent and discriminant validity issues. Proceeding to test hypotheses on the basis of this flawed scale with no proven construct validity was not correct.

Butler (1991)

This paper, which is actually a compilation of a series of studies, is aimed at developing a Conditions of Trust Inventory, as the author felt that other previously developed trust inventories dealing with a global measure of trust were not comprehensive enough to measure *conditions of trust*, a new aspect of trust measurement, for which theory needed to be developed, and research needed to be conducted.

A rational-theoretical approach was taken. The actual development of the scales was preceded by an excellent theoretical review, a clear definition of the core construct being measured, and a review of existing studies and their shortcomings in terms of psychometric properties. An initial study which involved interviews being conducted with a group of managers generated 454 conditions of trust and mistrust. These were classified into 10 categories on the basis of existing theory. A total of 11 scales were formulated from these items (one scale measured overall trust). The scales were refined using an iterative procedure using a total of 1531 management students in a set of 9 separate tests. The scales were then ready for assessment of homogeneity, reliability, and validity with the use of 7 different samples involving managers, subordinates, machine operators, and management students.

Jackson's four principles for scale construction and validation (1984) were explained, and the principles adhered to as closely as possible in the validation procedure. Based on results from initial studies, a theory of conditions of trust was developed, and a nomological net established. This was done to establish content and construct validity. Then test-retest, factorial homogeneity, and internal consistency reliability studies were carried out over a period of 6 years – these were shown to be good. Discriminant and convergent validity were established next with the use of a number of different trait measures, and a variety of methods. Good results were obtained. Construct validation was also conducted by using role playing tasks and observer ratings, and vertical dyads by generating hypotheses and testing them. The scales were shown to have good predictive validity also. The author concluded that the scales were psychometrically sound and could be used for all types of populations.

The outstanding strength of this paper is that the development and evaluation of the scales was conducted over a long period of time (several years); this gave plenty of time to establish and confirm reliability and validity. Another strength is the following of Jackson's principles, and the sequential system of scale development suggested by him. Many different situations and methods were used in a number of studies that helped establish convergent and discriminant validity. However, in spite of all the careful following of procedures, one fails to note any item level analysis, such as content saturation, and consequent validation of each item. Most of the analyses were conducted at the scale level. Also, no attempt was made to control for response style variance. It was argued that excluding items with social desirability content would remove relevant variance from trust. But no attempt was made to control for acquiescence or non-purposeful responding. This would have contributed to inflated reliability and validity values. A definite strength was establishing predictive validity by testing hypotheses. Overall, this was an excellent, detailed study that had a clear focus spanning the range of studies, and a sound procedural base. Analyses were not done separately for males and females, and also, it must be mentioned that generating the items initially with a group of managers might have precluded conditions of trust in other non-organizational relationships. Also, since the purpose of the entire study was also to develop theory in this area, a purely theoretical approach was not conducted.

Hargrave, Jennings, Anderson (1991)

The purpose of this paper is to construct a new scale that would help in developing an emerging theory, contextual theory, whose essence is the healing of human relationships through commitment and trust. According to this theory, there are 4 dimensions of relational reality that must be considered in therapy, and the purpose of this study is to develop a scale to measure one of them, *relational ethics*.

Since this was an emerging field, not much was available by way of theory, but the little available research that had been done was described, and a workable definition of the construct to be measured, provided. A rational approach was taken in scale development. Statements were first generated by the authors which reflected the various content of relational ethics according to the definitions developed. The 71 statements relating to “vertical” relationships and 65 “horizontal” relationships statements so generated were then evaluated by a panel of experts for face validity. Approximately half the items were retained. In the next stage, the preliminary Relational Ethics Scale was tested with a heterogeneous group of volunteer subjects. Separate analyses were conducted for different marital statuses, and two age groups, and differences noted. Item validity was good for both subscales. The items were then subject to principal components analysis; three factors were observed on each subscale. These were labeled. Items were seen to clearly load on the separated subscales, and good construct validity was inferred. High internal consistency values were observed. However, the vertical scale items were more clearly delineated than the horizontal scale items, and the same parallel was observed in existing literature on the topic. The finalized scale was then administered to other groups of volunteer subjects in two separate tests, in order to ascertain predictive validity, test hypotheses, and compare with results on other tests. Good results were obtained on all counts.

Since this was a developing academic field, and not much established theory was available, the right approach to take was the rational one. This was done in largely a systematic manner. Good definitions were provided and the item pool was generated based on these definitions. A great deal of care was taken in establishing face validity of the items. This was a definite strength, together with the item-level validation analysis they carried out. They tested separately for the marital groups and age groups, and this approach is superior to using a homogeneous group of subjects such as students. There is likely to be better generalizability. Reliability analysis was good, but more work needs to be done in terms of test-retest reliability, as this issue was overlooked. Item refinement was also conducted well, and preliminary convergent and discriminant validation tests were also good, but obviously insufficient – much more needs to be done in these areas.

No attempt was made to control for response style variance such as social desirability in any of the studies. This could have inflated reliability and validity values.

A number of different scales, such as the Dyadic Adjustment Scale, the Personal Authority in the Family System questionnaire, etc., were used in the final stage in order to establish convergent and discriminant validity, but the sample size here was very small (n = 36), when compared to the number of tests administered. A final critical comment relates to the development of the item pool. The authors themselves acknowledged that relational ethics is a complex construct, and that vertical and horizontal relationships might only be a portion of this construct. The scale developed might not be representative of all the complexities involved in this construct and this would be a weakness, especially if the scales were designed for wide usage in further research.

McCauley and Kuhnert (1992)

The purpose of this study was two-fold – to clarify the concept of *employee trust* in management in order to provide a framework for this and future studies, and to develop relevant scales and to do some exploratory work using these scales. The focus of this study was *organization-wide variables* and how they relate to employee trust, as opposed to earlier work, which concentrated on *job/ relational variables*.

The first half of the paper was devoted to an extensive literature review describing historical development of the concept being measured, a description of earlier empirical work and their results, and a detailed conceptual backdrop for the approach taken in this paper. Based on this, three sets of organization-wide variables were isolated as possible antecedents of trust – *professional development, job security,* and *perceptions of the organization's performance appraisal system.*

First, hypotheses were developed that related the above three sets of variables to the dependent variable, trust. Specific measures for each of these 4 variables were then constructed by borrowing items from several other existing scales and generating some, and combining these two. At this stage, definitions were provided. The composite scales were then administered as a voluntary survey in a large federal government training organization. First and foremost after results were in, reliabilities were calculated and found to be high and acceptable. This was taken by the researchers to imply good usability for the scales. A hierarchical regression was then conducted, with one set of variables being entered at each state, and being regressed against trust. Incremental R^2 was calculated at each stage. Then, a separate full model multiple regression was run, and the nature of this stage of analysis was exploratory. Results were analyzed in the light of the stated hypotheses and it was concluded that hypotheses were supported, and the scales developed could be used in further research in this area.

The strengths of this study are few and are far outweighed by the long list of weaknesses and potential flaws. An "upside-down" approach to developing scales was taken here. Firstly, after good conceptual development and background, which was a major strength of this paper, hypotheses were formulated. Then, scales were developed. This might have been acceptable had they developed their items from their theory. There is no evidence of this – items were selected and included on a rational basis, and there was no attempt to develop a comprehensive item pool first. The researchers appeared to have no idea of what is involved in scale construction apart from the simple reliability analysis conducted. There was no attempt whatsoever to validate the scales they developed. It is true that some of the items came from previously validated scales, but the new scales they generated this way should have been validated before use. A minor good point of the way they designed the scales was their awareness of the need to control for response-style variance – they negatively worded half the items in order to do this. Hypothesis-testing was good because they had sound theory, but this cannot be done until the measures have been validated in some way. There was no reason why after using hierarchical regression was used to test hypotheses, the researchers then went to doing an exploratory study using the full model multiple regression. Overall evaluation is that this is a very poor study, and the scales developed should certainly *not* be used in further research.

Strutton (1993)

A major portion of this paper involves constructing a new scale to measure psychological climate within an organization. The researchers were involved in work relating to organizational trust, and recent work in this area had generated a new construct, *psychological climate,* with which they wished to do more work.

They first began by giving a detailed definition of psychological climate, and a theoretical explanation of how this construct was related to that of trust in an organization. After explaining from theory how psychological climate was a multidimensional construct, the authors explicated the 7 hypothesized dimensions, defining each one. However, after doing this, they rationally derived a 28-item scale, with items covering 6 of the hypothesized dimensions, presented with a Likert-type format. For the 7th, which was trust, a previously validated measure, aimed at measuring just the concept of trust, was used.

460 randomly sampled sales organizations were requested to complete the two scales sent out. Only one salesperson from each organization was required to complete the items. 208 responses were received. The results were first analyzed by demographic groups such as sex, age and marital status to detect any systematic differences. None were found. The authors then proceeded to analyze the total sample together. A confirmatory factor analysis of the Psychological Climate scales was conducted, and the items were found to load on the 6 hypothesized dimensions, with high loadings. Two items that did not were dropped. Reliability analysis indicated high overall as well as separate internal consistency values for the scale and the sub-scales. The scale was then deemed acceptable, and was then used for testing some hypotheses relating to psychological climate and other constructs.

The strength of this paper was the good theoretical background provided up front. Clear definitions, distinctions, and relationships of the dimensions were provided. However, instead of capitalizing on this strength, and developing theoretically-derived items, the authors proceeded to rationally derive items for their scale, on the basis of what they felt were relevant for each dimension. There is no guarantee that these items covered the domain of items possible under each dimension. Just the fact that the scale exhibited high internal consistency values does not also guarantee the reliability; this could have occurred because of high redundancy of the items or their similarity. Test-retest reliability should have also been checked.

A procedural strength of their analysis was the exploration of possible differences on the basis of demographic groups. Other strengths were the use of confirmatory factor analysis to confirm the nature of the underlying structure, and the testing of hypotheses. However, these do not make up for the entire lack of any kind of validation of the measure. No attempt was made to establish convergent or discriminant validity, and this should have been done before using the scales were used for further hypothesis testing. Also, no controls were built in for response style variance, such as social desirability. Even though reliability values were provided for the subscales and deemed acceptable, no explanation was provided as to why values for only three of the subscales were given. Also, no item level analysis was conducted.

Considering the strong theory provided, the authors could have done a good job of developing sound theoretically-derived scales. The validity of the scale they developed is highly questionable, and its subsequent use, inappropriate.

Currall & Judge (1995)

The past few years have seen an explosion of writings and research on organizational and interpersonal trust. However, measurement of trust had been largely neglected. This study was motivated by the lack of a comprehensive measure of trust for use in organizations.

The paper begins with an excellent detailed theoretical review, and a clear definition of the trust construct. This was followed by an expression of the importance of construct validation right from the beginning stages of scale construction. A nomological net was established relating trust to other theoretical constructs. Then the 4 dimensions of trust were explicated, and several hypotheses generated, which would help establish the predictive validity of the measure.

In order to generate items, a preliminary set of interviews were conducted with a group of subjects and 26 items that described trusting behaviors, and that spanned the 4 theoretical dimensions were formulated. Then, the items were refined by testing on another 2 groups of subjects, resulting in a final scale of 20 items, with 5 on each of the dimensions. The scale was then ready for administration on another sample. The final sample consisted of 309 superintendents and 303 presidents of the National Education Association, and the American Federation of Teachers, a suitable sample as the researchers were investigating trust in interorganizational boundaries (here, the boundary was between the district's administration, and the local teachers union). 91% of the respondents were male, this figure representative of the population, and non-response bias was also checked for. 10 different analyses were carried out, all of them focused on establishing factor structure, convergent and discriminant validity, and adequacy of the nomological network, with the use of confirmatory factor analysis using LISREL. These 10 analyses are not described here, but one is used as an illustrative example. Fit of the 4-dimensional model was compared with a 3, 2, and 1-dimensional model; best fit was obtained only for the 4-dimensional model. This was taken as an indication of discriminant validity. Item-level and subscale correlations were also analyzed, helping to establish content saturation of the items. Hypotheses were tested, establishing predictive validity.

They also tried to establish the generalizability of the scale by repeating the analyses on other samples from the same population. They concluded that the scale had excellent construct validity and generalizability, and that it could be used for similar purposes in other organizational settings.

This study had numerous strengths. The researchers had a clear purpose and the whole study, comprising of several extensive surveys, was aimed at achieving this purpose. There was excellent theory development, and an added value was the nomological net. The pretests helped generate items, though not much is mentioned about concern about the adequacy of the item pool. Content saturation at item level was tested for, and this was done very thoroughly. Another strength is the realization that convergent and discriminant validity are factors that need to be looked into throughout the process of scale construction. The authors tried to incorporate this in every step of their procedures.

Some weaknesses that can be highlighted are as follows. No mention was made of the reliability of the measures (internal consistency, or test-retest). Secondly, it is necessary to examine correlations using external criteria also, with the use of different traits and different methods, for example, with a multitrait multimethod matrix, in order to concretely establish convergent and discriminant validity. Thirdly, the use of this particular sample, though more than adequate for this study, is a matter of concern. A limitation of the samples used in this study was that they were primarily male (over 90%). If the scale is to be used for other populations, its generalizability needs to be proven with repeated testing on different types of samples. Finally, response style variance was not controlled for, and this could have affected the reliability of the measures. Overall, an excellent study, but would have been made better had they followed a systematic method of scale development such as Jackson's (1970) sequential method.

Rotenberg and Morgan (1995)

This was a study to develop a scale to assess individual differences in children's ascription to the *trust-value basis for friendship.* The need for this scale was felt because no such measures existed, and such a scale was needed to further the research in the relevant area.

The study was preceded by a brief review of existing research in this area, and a definition of the construct provided. The studies and the review were largely empirical, and not much theory seemed to be available to help establish constructs and their relationships. The scale was intended to measure both *friendship preferences* as well as *actual friendships.*

Corresponding to these, two versions with 12 items each were developed. Items were adapted from other scales or added by the authors. Together with this True-Value Friendship (T-V-F) Scale, a previously validated Chumship Checklist was also administered to help establish discriminant and convergent validity. Subjects were 130 children from 5th and 6th grades in three Ontario Catholic schools, and the tests were administered in two sessions. Average ratings for friendship preferences and for actual friendships were subjected to an exploratory factor analysis with oblique rotation. Each facet revealed 3 clear factors, based on eigenvalue greater than one criterion. Correlations between the two scales administered were compared for convergent and discriminant validation, and test-retest as well as internal consistency reliability were calculated. Correlations were in the hypothesized directions and reliability was good. The researchers decided to 3 construct subscales for each facet based on the factor analysis results, and good reliability values were obtained for all except one subscale.

The study had few strengths and many drawbacks. One of the strengths of this study was the way the authors tried to build in discriminant validation right from the beginning. Internal consistency, test-retest reliability, and preliminary validation were good. A rational approach was used to generate items even though there appeared to be enough theory to be able to generate hypotheses. A comprehensive method to test convergent and discriminant validity could have been used, and more work needs to be done in this area, by using more scales in parallel. The subjects used in this study were very homogeneous (all students from Catholic schools), and this could be a problem for generalizing the results to children from varied school backgrounds. The basis on which items were generated was not adequate, and no item pool was used. The comprehensiveness of the 2 12 item measures is questionable. No pretests were conducted, and even a superficial face validity check was not conducted. A more detailed item-level analysis than item-factor loadings would have strengthened the validity of each item. Even though one subscale was weak, it was retained in the scale. Hypothesis testing was good, but the results are highly questionable when there was so little demonstration of its validity.

Couch, Adams and Jones, (1996)

From theory and prior research, two distinct conceptualizations of trust had been established – global trust, and relationship or relational trust. This study hypothesized a third type, network trust. As these 3 represent related but not identical constructs, this study was aimed at formulating a new scale to measure all three, and to explore the relationships between these three constructs and others with a number of hypotheses.

83 items were rationally derived and these were classified into 3 groups based on the 3 types of trust. The items were then administered to a group of undergraduates, and the scale's reliability explored. Items that performed poorly on this analysis were dropped. 50 items were left in the scale. Internal consistency and test-retest reliability were found to be good. The next study involved examining the convergent validity of the scale by comparing results with those obtained by 7 other existing measures of trust. Good convergent validity was seen. Construct validity was studied by conducting a principal components analysis with oblique rotation – 3 factors were extracted, and these corresponded with the hypothesized conceptualizations. 8 of the 50 items loaded on factors other than the subscale they were written for. Subsequent analyses on separate samples involved discriminant validity, studying the relationship of these 3 constructs with other constructs (basically locating them in the conceptual space within the interpersonal circumplex), and testing hypothesized relationships. Numerous other personality scales were administered for this purpose. Strong support was found for global trust, and relational trust, but not for a conceptually distinct network trust. The researchers concluded that their Trust Inventory was suitable for measuring the first two trust constructs, and called for additional research into the network trust construct.

This study is essentially a good one given some of the drawbacks of using the rational approach in scale construction. A very thorough investigation of reliability of the scales, discriminant, and convergent validity was carried out, though it must be pointed out that, instead of separate analyses, these three issues could have been explored simultaneously with the help of a multitrait-multimethod analysis. Many different measures were used, but all methods were the same. A very large sample was available to them (N = 1229), and this kind of analysis could easily have been carried out. The rational strategy was the right one to use in this case given the lack of theoretical support for the network trust construct.

Some of the weaknesses in the study are now discussed. Despite mounting evidence against the third trust construct (network trust), the authors retained these items in the scale, and called for more research. There was enough evidence against such a construct, and it should have been abandoned. For example, in the principal components analysis, the third factor, network trust, accounted for only about 5% of total variance. No controls were made for response style variance, and this is a weakness of all rational-based approaches. Separate analyses should really have been conducted for males and females considering research that indicated that males and females display different trust profiles. Also, predictive validity needs to be established together with more extensive work on discriminant and convergent validity.

Cummings and Bromley (1996)

The purpose of this paper is to present the conceptual and empirical development of a measure of organizational trust. The conceptual background for the approach taken in developing the scale was grounded in transaction cost economics, and a formal, explicit, multidimensional definition of trust was developed and provided from this source.

The definition of trust included 3 dimensions, and the authors expressed the need to measure trust across three components – as an affective state, as a cognition, and as an intended behavior. Survey items were to be constructed for each of the three dimensions of trust. A group of 5 doctoral students generated 273 items, and after an evaluation of face validity, 121 items were retained. Another 15 items were added from the previously validated Organizational Commitment Questionnaire, in order to establish discriminant validity. Inter-rater agreement on which dimension each item belonged to was determined. Then, the researchers went through each item and dropped what they thought were redundant items so as to reduce the questionnaire length.

The final survey was then administered to a group of 323 employees, students, and MBA students of a University. Latent variable structural equation modeling was used to confirm the underlying 3 dimensions, and to assess reliability and validity. The model fit was good, item-factor correlations were generally high (over 0.60), reliability of each dimension was high, and good discriminant validity was observed between the two scales used. Some items, particularly the behavioral intent items did not have high item-factor correlations. As a final analysis, a predictive study was conducted for testing hypotheses, and good results were obtained. The researchers also developed a short form of the Inventory and after validating it also, concluded that they had a good measure of organizational trust.

The approach taken here is a theoretical one, and thus has a strong foundation in established theory. The researchers developed their items based on the theory and definitions they generated – this was a definite strength of the paper and their approach. They knew the nature and the number of dimensions that were involved in the construct, and therefore they took a confirmatory rather than an exploratory approach in their use of structural equation modeling, in order to confirm the hypothesized structure. The item pool had good face validity and their preliminary exploration of reliability and validity was good.

However, it must be emphasized that this was only preliminary. The work that was done here is not sufficient. Test-retest reliability, and further work on discriminant and convergent validation needs to be done, for example with the use of the multitrait-multimethod analysis, especially as other measures of trust and other related constructs were available. A flaw was their dropping of items by just examining the items for redundancy – a better way would have been to combine this with an analysis of the item-level correlations. Another potential weakness was their use of *student* subjects for developing an *organizational* trust inventory. Separate analyses should have been done for employees and students, and for men and women. No control was also made for response style variance. Overall, this is a good study and a good development of the scale, but it would be strengthened by following the improvements suggested above.

Nyhan and Marlowe Jr., (1997)

This article reports on a study conducted to develop a 12-item scale to measure an individual's level of trust in his or her supervisor and in his or her organization as a whole (the Organization Trust Inventory (OTI). The researchers were motivated to construct this scale because they felt that existing measures of organizational trust were limited in scope. They used a theoretical approach.

Initially, the organizational trust construct was defined, and its relationship with other constructs explicated, by means of existing theory. The usefulness of measuring trust in organizations was explained, and a suitable theoretical framework for developing the construct was arrived at; this was the Luhman’s two-dimensional theory of trust.

After a review of literature, a 12-item 7-point Likert-type scale was constructed, with 8 items measuring trust in supervisors and 4 items measuring trust in the organization as a whole. Then, some pretests were conducted on four small, primarily male, groups to establish reliability. High reliability values were found. The researchers subsequently conducted an exploratory factor analysis and found 2 components that corresponded with the two theorized components. Once this was done, the researchers went ahead and tested the scales on a larger sample of employees of a county government planning department. This sample had better representation in terms of the sexes. Internal consistency and test-retest reliability (after 4 days) were computed, and were found to be high. Confirmatory factor analysis was conducted, and the hypothesized two-dimensional structure was obtained. Convergent and discriminant validity was explored by comparing results from the pre-tests with those from selected other tests, and hypothesis testing. Results were in the expected directions.

The authors concluded that the results of the studies demonstrated that the OTI was psychometrically adequate and stable and could be used as a reliable and valid measure in an organizational or research setting.

The study described here is a very thorough one designed to construct and evaluate the OTI scale. Overall, it is a good study, with many strengths, such as a good theoretical exploration, reliability and validation studies. However, a few flaws mar the perfection that was aimed at, and these are highlighted here. Firstly, the development of the item pool was not described, so there is no way for readers to judge the adequacy of it.

Secondly, response style variance was not controlled for, so this could have inflated correlations. Subjects used in the preliminary studies were primarily male, and this could have affected results, as there is evidence from other research that profiles of trust are very different for males and females. Also, there was really no need to carry out exploratory factor analysis in the pre-test stage – theoretical evidence was very strong for a two-dimensional trust construct. This would have sufficed to explicate the structure; at the most, a confirmatory factor analysis could have been carried out. Convergent and discriminant validation could have been established by means of the multitrait multimethod analysis, and this could have saved a lot of time and effort, when compared to the analyses actually done by them. Another small point is that test-retest reliability was conducted with a gap of four days; this is not sufficient. At least 3 weeks' interval is normally recommended. These minor improvements would go a long way to further establishing the usefulness of the test.

McAllister (1998)

This study addresses the nature and functioning of relationships of interpersonal trust among managers and professionals in organizations, the factors influencing trust's development, and the implications of trust for behavior and performance. Substantial quantities of theoretical support were available from prior research, but since the current author was introducing two new subconstructs – *cognitive*, and *affect-based trust*, new measures needed to be developed focusing on these two.

Substantial theoretical background was provided, and a theoretical model and corresponding hypotheses were generated based on this theory. Definitions were provided differentiating these two subconstructs. Predictions were also made relating to how these would affect behavioral outcomes.

A 48-item pool was created from a review of literature and existing measures of trust. A panel of experts was then asked to review each item in order to provide face validation. Items that were poor or which were ambiguous were dropped. 20 items remained, one for each type of trust. An exploratory factor analysis based on pretest results further reduced the number of items to 11 of the strongest loading items. A 25-item measure of behavioral response was also developed in the same way. A sample of 194 managers and professionals were selected, and the test administered by way of a questionnaire.

Confirmatory factor analysis was then conducted using LISREL. Reliability estimates were high, and the model fit the data well. Discriminant validity was tested by constraining a single phi coefficient. There was a significant worsening of the model, showing good discrimination between different constructs. Some behavioral subconstructs did not show good discrimination, and these were dropped. Systematic tests for group differences were made (eg. males versus females). Hypotheses were tested using a nested-models approach. Overall, the measure proved very adequate for the purpose, and hypotheses testing supported the two new subconstructs.

The strength of this paper is the strong theory support provided throughout – it provided a clear focus at every stage of the study. Hypotheses and items were generated on the basis of the theory. However, a 48-item pool may not be entirely representative of the domain of items representing the two constructs. Another distinguishing feature was the use of subjects appropriate for the purpose of the study – organizational subjects were used, but the limitation to only managerial subjects draws into question whether the nature of trust could be different for other groups, for example, blue-collar workers. A confirmatory rather than exploratory approach was used, as there was testing of theory-based hypotheses. Separate analyses were conducted for each sub-group, and the results compared for differences. Face validation of the items, reliability analyses (though test-retest reliability was not assessed), and initial discriminant validation was provided. Also, the testing of hypotheses tried to establish some predictive validation for the scales. Although item-factor correlations were considered, a more detailed content saturation analysis should have been carried out. Inflated reliability and validity values could have been obtained because acquiescence or social desirability was not controlled for. Items should have been balanced in terms of true-keyed and false-keyed ones. Finally, more work needs to be done on construct validation, by using the scales with other related and unrelated measures and comparing their performance.

Murstein, Wadlin, and Bond (1987)

The purpose of this study was to formulate a revised version of the Exchange Orientation Scale because of severe limitations exhibited by the earlier version, such as the purely subjective nature of choice of items for inclusion by the earlier researchers.

Initially, a brief review of the literature pertaining to *exchange-orientation in relationship to marital adjustment* was provided. This included a discussion of the relationship between this and other related and relevant constructs. 56 items were generated; these were drawn from the earlier test and combined with some other generated by the authors based on what they felt were relevant. A group of 61 college students were then asked to judge and rate how closely each item reflected the "exchange" concept, as defined by the authors in the test. Based on this analysis, 21 items having the highest means (that is, those judged to have the closest relation to exchange as defined) were retained.

These items were then administered to 32 volunteer married couples along with a battery of unrelated tests. Only 2 items were reverse-scored to avoid confusion observed in earlier studies. Analysis was conducted separately for men and women. Internal consistency was calculated and found to be good. Discriminability was examined by comparing husbands in the upper quartile with husbands in the lower quartile for each item, and significance of the difference tested. The same was done for wives. Good discriminability was observed for all items for wives, and all but two items for husbands. These were dropped. The researchers concluded that the present version of the scale was more thoroughly constructed and more justifiable than the earlier version.

The single major contribution of this study was the derivation of separate scales for men and women and the separate analyses done that enabled their formulation. Apart from this, the item-level analysis was good, although it must be pointed out that there were really too few subjects (32 in each group), to justify this. The judgment of face validity was a good idea, but perhaps should have been done with a panel of experts in the field also. The way the authors generated items for this study was subject to the same bias that they accused the other set of researchers of having in the earlier study – a great deal of subjectivity was used, rather than theory. Also, the study would have added much more value if the definition of numerous constructs and their relations provided in the beginning could have led to generation and testing of hypotheses relating to them. Reliability analysis was good, but the scale needs test-retest reliability, and convergent and discriminant validation, of which there is virtually no evidence. Overall, the value added by this study is questionable.

Hargrave and Bomba (1993)

The main purpose of this scale was the further validation of the Relational Ethics Scale (the RES) constructed and tested by Hargrave et al., (1991), as there was a need expressed to further validate it with clinical and nonclinical populations. Also, since different results were obtained for different marital statuses, and age, more research into these issues were called for.

In study 1, the RES was administered to a homogeneous group of single, never-married undergraduate volunteers to determine its reliability and validity. Principal components analysis was conducted with the results, and item level analyses were done. Good item discrimination was demonstrated between the top and bottom quartiles of scores, and reliability value was good, but lower than in the previous study.

Principal components analysis revealed 3 components for both horizontal and vertical statements. Some items did not load in the same way as in the original study. The trust and justice components accounted for the major portion of variance in both subscales, just as in the earlier study. Scores from the present study were compared with different subgroup scores in the earlier study. Significant differences were observed on the basis of marital status and age, but comparable results were observed between the same marital status and age groups in both studies. The authors concluded that the RES was valid and reliable among single, never married individuals. Stronger item and construct validity was observed for the horizontal rather than the vertical subscale. Study two was conducted to explore the differences in marital status and age groups observed in study one, but the subjects and results were not described in detail here. Further research was recommended.

The purpose of this study was clearly the further validation of the results obtained in the earlier study. When this was the case, a more heterogeneous subject-group should have been used, similar in composition to that in the earlier study. It was incorrect to compare results in this study directly with the subgroups in the earlier study. A strength was the item level analysis conducted, for which significant differences with each and every other item were examined using t-tests. This would enhance the content saturation of each item. When the results of the earlier study had revealed 3 components in each subscale, a confirmatory rather than an exploratory approach should have been used in this study. Even though the need was expressed to test on clinical and nonclinical samples, this was not done. Also, some of the careful following of procedures observed in the previous study and preliminary discriminant validation done were not reproduced here, even though the purpose of this study was further validation. In sum, there is very little by way of added value contributed by this study. What would have really helped would have been a further exploration into discriminant and convergent validity of the scales by using more number of scales measuring similar and differing constructs, and predictive validation by testing hypotheses. Separate studies using clinical versus nonclinical samples would have added value.

Couch and Jones (1997)

The purpose of the current research study was to further validate the Trust Inventory constructed by Couch et al., (1996). This was an innovative self-report measure that partitioned trust into separate domains including Partner Trust, Network Trust, and Generalized Trust. A second purpose was the testing of several hypotheses related to the trust construct, as part of validation, and as part of extending earlier research.

Initially, a brief review of recent developments in theory as well as a clear definition of the three domains of trust was provided. Subjects involved in the study were a large group of undergraduate psychology students who were involved in a romantic relationship for a mean period of 22.9 months. Each subject completed the Trust Inventory as well as at least one other measure. These measures were either alternative measures of global and relational trust, or a selection of measures of various personality, emotional, and relationship constructs.

Initial reliability estimates exceeded those found in the earlier study. Interitem correlations were deemed acceptable. Test-retest reliability over a 9-week period indicated temporal stability. Then, each subscale was analyzed with numerous other scales measuring closely-related, as well as those measuring distinct, different constructs, in order to obtain estimates of convergent and discriminant validity. For example, partner trust, as measured by the Trust Inventory subscale was compared with the Faith subscale of the Trust Scale, the Emotional Trust subscale of the Specific Interpersonal Trust Scale, and the Dyadic Trust Scale. It was also compared with scales measuring theoretically unrelated constructs The reliability of each subscale was good. For each subscale, predictions were made about the direction and strength of relationship with other scales. The results indicated good convergent and discriminant validity for two of the subscales – Partner and Generalized Trust. The extent and direction of correlations were generally according to prediction. Values for convergent validity for Network Trust were not calculated because it was a new construct for which there were no comparable measures. It did demonstrate moderate discriminant validity. It was generally concluded that there was good evidence of validity and reliability of the Trust Inventory.

This is an excellent study that explores the issue of convergent and discriminant validity and reliability in a very thorough, systematic way. Many different measures were used to test convergent and discriminant validity, but different methods were not used (all the measures were self-report). Predictions of relationships of the various constructs were made based on theory and these were done *a priori.* Both test-retest as well as internal consistency reliability were established. The only comment that can be made relating to this analysis is that using a Multitrait-Multimethod matrix analysis would have made the study even more thorough and systematic. Another good point was that analyses were made separately for men and women, that helped them identify sex differences. Social desirability and other types of response style variance were not controlled for, and this might have contributed to very high reliability values. If this study were to be repeated, the one thing that should be changed is the addition of different methods of measuring the same traits. This would make the validation more tight.

General Overview and Conclusions

Regarding the strategy used in scale construction, 8 studies used the Rational approach, 5 used a purely Theoretical strategy, and the remaining used a combination of the two methods. There were no instances of the Empirical strategy being used. We gave an overall rating (last column in summary table) to each study based on appropriate use of strategy, correct procedures, and overall execution of the study as well as its discussion and write-up. 4 (20%) were rated “Excellent”, 2 had “Very Good” rating, 7 (35%) were rated “Good” and the remaining 7 were rated “Poor”, or “Very Poor”. A large proportion of studies appears in the “Poor” category, and a more detailed analysis of why this was so follows in a later part of this discussion. Looking at the issue of subjects used, there is a wide cross-section of the types of subjects used, from college undergraduates, and heterogeneous volunteers, to organizational workers, managers, and federal government employees. Almost all (19) had a substantial sample size to work with, and this is important, because of the nature of the analyses that accompany scale construction.

A general observation is that not one study used the Campbell and Fiske criteria or the Multitrait Multimethod matrix in analyzing convergent and discriminant validity. It is hard to understand why when the M-M matrix provides a very comprehensive and systematic way of evaluating validity. It can be understood from pioneering studies in an area, measuring constructs for the first time, that there will not be enough evidence, from a variety of methods, to use this technique. However, there is no reason why with later studies, particularly those conducted in the 1980’s and 1990’s, when the construct of trust was fairly well-established, this was not done.

A few studies had a very good theoretical portion and description but only one used Jackson’s criteria explicitly, attempting to follow the four principles laid out by him.

Most of them, whatever the strategy they used, seemed to realize that the starting point for a good scale construction was a clear definition of the construct to be measured. Even the ones rated "Poor" did this, although there were a handful that failed to do so. Some studies were weak because they used a rational strategy approach even though there was good theory available. According to Jackson (1970), there can be no substitute for good theory to guide test construction. In the absence of theory, a rational approach may be used, but such scales have to be validated very carefully. A number of rationally-derived scales (at least 8 of them) were formulated without generating an item pool initially. From this sample, it appears that many researchers do not seem to be aware of the importance of generating as comprehensive an item pool as possible before refining or deleting items. Even if a rational strategy is used, the scale cannot be generalizable if the domain of items has not been covered.

Jackson also emphasized the need to control for response style tendencies, such as social desirability, random responding, and acquiescence which, if uncontrolled, could override hoped-for content consistencies, and can cause inflated scale correlations. At least 12 of the present studies made no attempt to do this, and the subject seemed to be one on which there was a lot of ignorance. Only one study (Larzelere and Huston, 1980) controlled for this, and went ahead and did item level analysis based on this, with an index of content saturation such as the Differential Reliability Index.

The need for homogeneity of the scales seems to be an issue pretty well understood by most researchers developing or validating scales. Although all but one of them (Currall and Judge, 1995) tested for and reported internal consistency reliability, 8 of them did not bother with test-retest reliability. They could not possibly have concluded that reliability was good if they had not established this.

Convergent and discriminant validation, possibly the most important part of scale evaluation was weak in most cases. Few studies tried to build in these processes right from the beginning, while constructing the scale. Those that did, for example, Butler, 1991, were excellent. Many made some attempt to establish discriminant validity as part of their study, but stopped short of convergent validation, or made inappropriate conclusions about the adequacy of their validation studies. One cannot expect detailed validation from earlier studies in this field, because the area was still unexplored and needed to be developed, but more recent studies should pay more attention to the importance of construct validation.

We conclude with some general remarks on this area of research. Research on trust, as a measurable construct, began in the 1950's and 60's when Rotter's (1967) Interpersonal Scale was constructed. It was an excellent study and remains to this day an important, much used scale, and often-cited paper. With subsequent developments in the research into this construct, researchers, instead of building on the strong foundations provided by the earlier researchers, tended to come out with narrow, often weak, formulations of scales of their own. Instead of adding value, the net contribution of such scales appears to be more confusion instead of clarification. In recent years, the trend appears to have changed, with many more researchers focusing on building on and further validation of earlier work. This is a good trend that can only benefit the field. However, if many more researchers went back to the basics of good scale construction, and followed the principles laid out by pioneers in this field, there is potential for much more benefit to the researchers, and subsequent progress in research in this area. More work also needs to be done on developing different methods of measuring trust, not just self-report measures.

Appendix A - Summary Table of Review Papers

(See Excel File)

References:

1. Butler, John K., (1991), *Toward Understanding and measuring conditions of trust: evolution of a conditions of trust inventory,* Journal of Management, 17, 3, 643-663.
2. Clark, Margaret S., Robert Ouellette, Martha C. Powell, and Sandra Milberg, (1987), *Recipient's Mood, Relationship Type, and Helping,* Journal of Personality and Social Psychology, 53, 1, 94-103.
3. Cook, John & Toby Wall, (1980), *New work attitude measures of trust, organizational commitment and personal need non-fulfillment*, Journal of Occupational Psychology, 53, 39-52.
4. Couch, Laurie L., and Warren H. Jones, (1997), *Measuring Levels of Trust,* Journal of Research in Personality, 31, 319-336.
5. Couch, Laurie, L., Jeffrey M. Adams, and Warren H. Jones, (1996), *The Assessment of trust orientation*, Journal of Personality Assessment, 67(2), 305-323.
6. Cummings, L.L., and Philip Bromley, (1996), *The Organizational Trust Inventory*, in Kramer, Roderick M., and Tom R. Tyler, (eds.), Trust in Organizations, Sage Publications, Thousand Oaks, CA.
7. Currall, Steven C., and Timothy A. Judge, (1995), *Measuring Trust between organizational boundary role persons,* Organizational Behavior and Human Decision Processes, 64, 2, 151-170.
8. Hargrave, Terry D. and Anne K. Bomba, (1993), *Further validation of the Relational Ethics Scale,* Journal of Marital and Family Therapy, 19, 3, 292-299.
9. Hargrave, Terry D., Glen Jennings, & William Anderson, (1991), *The development of a relational ethics scale,* Journal of Marital and Family Therapy, 17, 2, 145-158
10. Johnson-George, Cynthia, & Walter C. Swap, (1982), *Measurement of specific interpersonal trust: Construction and Validation of a scale to assess trust in a specific other,"* Journal of Personality and Social Psychology, 43, 6, 1306-1317.
11. Lagace, Rosemary, and Gary K. Rhoads, (1988), *Evaluation of the Macdonald, Kessel, and Fuller Self-Report Trust Scale,* Psychological Reports, 63, 961-962.
12. Larzelere, Robert E., & Ted L. Huston, (1980), *The Dyadic Trust Scale: toward understanding interpersonal trust in close relationships*, Journal of Marriage and the Family, Aug., 595-604.
13. McAllister, Daniel J., (1998), *Affect and cognition-based trust as foundations for interepretation,* Academy of Management Journal, 38, 1, 24-
14. McCauley, Dan P., and Karl W. Kuhnert, (1992), *A theoretical review and empirical investigation of employee trust in management,* PAQ, Summer, 265-284.
15. Murstein, Bernard I., Robert Wadlin, and Charles F. Bond, (1987), *The Revised Exchange-Orientation Scale,* Small Group Behavior, 18, 2, 212-223.
16. Nyhan, Ronald C., & Herbert A. Marlowe, Jr., (1997), *Development and psychometric properties of the organizational trust inventory,* Evaluation Review, 21, 5, 614-635.
17. Rotenberg, Ken J., and Cathy Morgan, Jr., (1995), *Development of a scale to measure individual differences in chidlren's trust-value basis of friendship,* Journal of Genetic Psychology, 156, 4, 489-
18. Rotter, J.B., (1967), *A new scale for the measurement of interpersonal trust,* Journal of Personality, 35, 651-665.
19. Rubin, Zick, (1970), *Measurement of romantic love,* Journal of Personality and Social Psychology, 16, 2, 265-273.
20. Strutton, David, Al Toma, & Lou E. Pelton, (1993), *Relationship between psychological climate and trust between salespersons and their managers in sales organizations,* Psychological Reports, 72, 931-939.
21. Wrightsman Jr., Lawrence S., (1964), *Measurement of Philosophies of Human Nature,* Psychological Reports, 14, 743-751.

					Appendix A - Summary Table of Review Papers				
No.	**Authors**	**Year**	**Trait**	**Instruments**	**Subjects**	**Strategy**	**Strengths**	**Weaknesses**	**Overall Rating**
1	Wrightsman, Jr.	1964	Philosophies of human nature	Philosophies of Human Nature Scale	907 undergraduate and graduate students in 3 separate studies	Rational	Reliability, discriminant validation, no theory therefore rational approach appropriate, separate analyses for different groups, acquiescence, hypothesis testing	No item pool generation; items may not be comprehensive, total lack of validation except for preliminary discriminant validation, no factor analysis to see if items fell out in hypothesized dimensions	Good
2	Rotter	1967	Interpersonal trust	Rotter Interpersonal Trust Scale	248 male, 299 female UG psychology students	Theoretical	Theory laid out, controlled for social desirability, reliability, construct and discriminant validation done	Nomological net, item pool generation not described, could have done more validation and testing of hypotheses also	Excellent
3	Rubin	1970	Romantic love	Love and Liking Scale	198 undergraduates and 158 dating couples	Rational	Reliability, discriminant validation, no theory therefore rational approach appropriate, separate analyses for different groups, social desirability, hypothesis testing, item level analysis	Test-retest reliability, incorrect factor analysis, dropping items, more convergent validation needed	Good
4	Cook & Wall	1970	Interpersonal trust at work	Scale of Interpersonal trust at work	650 blue collar male workers, and volunteers	Theoretical-rational	Good definitions from theory, content saturation, test-retest, coefficient alpha, discriminant and preliminary convergent validation	All male subjects, could have used M-M method instead of just examining correlation matrix, more construct validity needed, no item pool, social desirability	Good
5	Larzelere & Huston	1980	Dyadic interpersonal trust	Dyadic Trust Scale	120 females & 75 males, young, involved in close relationships	Rational	Good theory development, response style variance controlled, DRI, content saturation,reliability, face validity, discriminant validity	Theoretical development not followed in item generation, test-retest reliability not done, convergent validity	Good
6	Johnson-George and Swap	1982	Interpersonal trust	Specific Interpersonal Trust Scale - Males and Females	180 male, 255 female UG psychology students	Rational	Discriminant validity; reliability analysis. Separate analysis for males and females	Small no. of items in item pool, no convergent validation, theoretical strategy should have been used, no control for response style variance, test-retest reliability not done	Poor
7	Clark, Ouellette, Powell, & Milberg	1987	Communal Orientation	Communal Orientation Scale	over 1000 undergraduate students	Rational	Few; control for response style variance, reliability analysis,	Questionable item pool; incorrect factor analysis; no convergent and discriminant validity	Poor
8	Butler	1991	Conditions of trust	Conditions of Trust Inventory	More than 2000 students, managers, subordinates and machine operators in a series of studies	Rational-theoretical	Study spanned several years - time to establish reliability, convergent, discriminant, and predictive validity, Jackson's principles followed, good theory development and building, nomological net, procedurally sound	No evidence of item level analysis, no control for response style variance, items might not be generalizable to all types of populations	Excellent
9	Hargrave, Jennings & Anderson	1991	Relational Ethics	Relational Ethics Scale	406 heterogeneous volunteers	Rational	Good item development, item level anlaysis, item refinement; reliability and preliminary convergent and discriminant validation, predictive validation, testing of hypotheses	Test-retest, no control for response-style variance, more needed on convergent and discriminant validation, not all dimensions of construct might have been covered	Very good
10	McCauley & Kuhnert	1992	Employee trust in management	Employee trust in management scale	293 federal governement employees	Theoretical-rational	Good theory, reliability, some attempt to control for response style variance	No item pool generation, items developed on a rational basis even though there was strong theory, no validation entirely, hypothesis testing with scales even though there was no validation	Very poor
11	Strutton, Toma & Pelton	1993	Psychological climate	Psychological Climate Inventory	208 salespeople from sales organizations	Theoretical-rational	Good theory, reliability, CFA, examined for systematic demographic differences, testing of hypotheses	Test-retest, comprehensiveness of item pool, no validation, no control for response-style variance, use of rational instead of theory-derived items	Poor

No.	Authors	Year	Trait	Instruments	Subjects	Strategy	Strengths	Weaknesses	Overall Rating
12	Currall & Judge	1995	Organizational trust	Trust Questionnaire	309 superintendents, 303 presidents of public school administration	Theoretical	Excellent theory, nomological net, hypotheses, discriminant and convergent validation using CFA, good item generation, but comprehensive?, content saturation at item level	Reliability analysis missing, generalizability as sample was 91% male, needs external validation for example with multitrait multimethod matrix.	Excellent
13	Rotenberg & Morgan	1995	Trust value basis for friendship preferences as well as actual friendship	Trust-Value Friendship Scale	130 children from Canadian Catholic schools	Rational	Few; reliability, a little discriminant validation, hypothesis testing	Item pool not generated, no face validity checks, no pretest, subjects too homogeneous - study or scale may not be generalizable, more validtaion, gender differences not tested.	Poor
14	Couch, Adams, & Jones	1996	Generalized, relational, and network interpersonal trust	Trust Inventory	1229 undergraduates	Rational	Good investigation of reliability, discriminant, and convergent validity	Could have used multitrait-multimethod analysis, network trust should have been dropped as there was no evidence to support it, predictive validity, separate analysis for men and women	Good
15	Cummings & Bromley	1996	Organizational trust	Organizational Trust Inventory	323 employees and MBA stuednts at a University	Theoretical	Good theoretical approach to item generation, structure known from theory, so used confirmatory analysis, good face validity, predictive validity, reliability and initial validation	Test-retest, response style variance, pretesting not done, item selection done by judgment, need further discriminant and convergent validation, validity of using students?	Good
16	Nyhan & Marlowe Jr.	1997	Organizational trust	Organizational Trust Inventory	95 male and 107 female employees	Theoretical	Theory laid out, reliability, construct, convergent and discriminant validation done	Item pool generation not described, subjects in pre-tests primarily male, EFA conducted when not needed, response style variance not controlled for.	Very good
17	McAllister	1998	Cognition- and Affect-based organizational trust	Trust and Behavior measure	194 managers and professionals	Theoretical	Strong theory guided all steps of the study, hypotheses generated, subjects relevant for measure, reliability, initial discriminant validation, predictive, and face validation	No item-level content saturation analyses, managerial trust may be different from trust in workers, test-retest, social desirability, balance true-keyed and false-keyed items, more work on convergent an discriminant validation needed	Good
				FURTHER	**VALIDATION**	**STUDIES**			
18	Murstein, Wadlin, and Bond	1987	Exchange orientation	Revised Exchange Orientation Scale	61 college students, 32 married couples	Nil	Separate analyses for men and women generated separate subscales, item level analysis, reliability, good review of theory	Item generation subjective, even though theory was available, small sample size for item analysis, hypothesis testing could have been done, no discriminant and convergent validation, test-retest, little vaue added	Poor
19	Hargrave and Bomba	1993	Relational Ethics	Relational Ethics Scale	162, single never-married undergraduate volunteers	Nil	Good item level analysis	Very little value added to earlier study; homogeneous subject group used, not comparable with earlier study, PCA inappropriate here, direct comparison with subjects in earlier group inappropriate, further validation and testing of hypotheses should have been done.	Poor
20	Couch and Jones	1997	Generalized, relational, and network interpersonal trust	Trust Inventory	445 undergraduate students involved in a romantic relationship	Nil	Very thorough and systematic exploration of convergent and discriminant validation, reliability, separate analyses for men and women, a priori specification of relationship with other constructs	Could have used multitrait-multimethod analysis, social desirability not controlled for.	Excellent