European Journal of Personality, Vol. 4, 209–235 (1990)

The Strelau Temperament Inventory-Revised (STI-R): Theoretical considerations and scale development

JAN STRELAU
University of Warsaw, Poland and University of Bielefeld, FRG

ALOIS ANGLEITNER and JÜRGEN BANTELMANN
University of Bielefeld, FRG

WILLIBALD RUCH
University of Düsseldorf, FRG

Abstract

The development of a revised Strelau Temperament Inventory (STI-R) is reported. It is assumed that the STI-R provides a measure of the basic central nervous system (CNS) properties (strength of excitation, strength of inhibition, and mobility of the CNS) as understood by Pavlov. On the basis of a series of studies, the development of the final forms of the revised STI has undergone several steps. The following forms have been elaborated: (1) a 252-item pilot form of the STI-R; (2) a 166-item STI-R with 'yes' and 'no' answer format; (3) a short form (84 items) of the STI-R (STI-RS) with 'yes' and 'no' answer format; (4) a 166-item STI-R with a 4-point Likert scale; and (5) an 84-item STI-RS with a 4-point rating scale. The psychometric characteristics of the consecutive versions of the revised STI improved from step to step, and in general these characteristics are judged as being satisfactory. Especially recommended by the authors are versions (4) and (5), which have, among other things, the highest reliability scores. They are regarded as the final forms of the STI-R and STI-RS.

INTRODUCTION

Pavlov's theory of the types of the central nervous system (CNS) introduced in the first quarter of the twentieth century has gained increased popularity in the last decade, especially among the biologically-oriented personality researchers (e.g. Buchsbaum, 1978; Claridge, 1985; Eysenck, 1972; Strelau, 1983; Zuckerman, 1979). The reason for the renewed interest in the properties of the central nervous system

Requests for reprints should be sent to: Jan Strelau, Faculty of Psychology, University of Warsaw, Stawki 5–7, 00–183 Warsaw, Poland. The address of Alois Angleitner is: Department of Psychology, University of Bielefeld, P.O. Box 8640, 4800 Bielefeld 1, FRG.

0890–2070/90/030209–27$13.50

Received 27 April 1989
Accepted 27 October 1989

may be explained by at least two facts. First, Pavlov's typology offers the most adequate physiological interpretation of the Hippocrates–Galen types of temperament, the latter still being popular among professionals and laymen. Second, the Pavlovian constructs of strength of the CNS and of protective inhibition are closely related to the concept of arousal (activation) to which most biologically-oriented personality theories refer (see Strelau and Eysenck, 1987).

Studies on CNS properties, especially popular in Eastern Europe (Nebylitsyn, 1972; Strelau, 1969, 1983; Teplov, 1964), are also conducted in the West, mainly with the aim to prove the construct validity of arousal-oriented personality dimensions, or to search for links between these dimensions and the Pavlovian traits of temperament. As examples, studies conducted on extraversion (Carlier, 1985; Loo, 1979; Stelmack, Kruidenier and Anthony, 1985), augmenting–reducing (Barnes, 1976; Kohn, Cowles and Lafreniere, 1987), or sensation-seeking (Goldman, Kohn and Hunt, 1983; Zuckerman, Kuhlman and Camac, 1988) may be mentioned here.

One of the main difficulties in studying the CNS properties consists in the lack of psychometric tools aimed at measuring the Pavlovian constructs. The only questionnaire which allows us to assess the behavioural correlates of the hypothetical, pseudophysiological constructs of strength of excitation (SE), strength of inhibition (SI), mobility (MO), and balance (BA) of the CNS properties as understood by Pavlov is the Strelau Temperament Inventory (STI), constructed by the first author at the end of the 1960s (Strelau, 1972). This questionnaire, translated into many languages, including Chinese and Japanese, gained some international popularity. In several studies (Carlier, 1985; Stelmack *et al.*, 1985), but especially on the basis of our own psychometric research (Strelau, Angleitner and Ruch, 1989), it has been shown that the STI, in spite of its rather satisfactory construct validity (Strelau, 1983), is lacking in many psychometric characteristics necessary to accept this diagnostic tool as a valid measure of the CNS properties. Our studies (Strelau *et al.*, 1989), conducted on four independent samples, comprising altogether over 800 subjects, allowed us to conclude, among other things, that the STI scales intercorrelate higher than expected theoretically (e.g. SE and SI up to 0.38, SE and MO up to 0.59); that they contain too many items; and that they are highly loaded by social desirability (e.g. both SE and SI correlate 0.43 with a Social Desirability scale developed in the German version of Jackson's PRF). Furthermore, many items show extreme endorsements, and the item–scale correlations are unsatisfactory (e.g. 19.4 per cent of the items from the MO scale have item–total correlations below 0.20).

The disadvantages of the original STI as well as the belief that the Pavlovian concepts of CNS properties are fruitful constructs to be applied in the research centred on biologically-based personality/temperament dimensions stimulated us to develop a new version of the STI. The research aimed at constructing the *Strelau Temperament Inventory-Revised* (STI-R) presented in this paper is guided by advanced psychometric personality scale construction strategies, as proposed by Angleitner, John and Löhr (1986).

THEORETICAL CONSIDERATIONS

Pavlov's concept of the CNS properties has been presented recently in several publications (Mangan, 1982; Strelau, 1983; Strelau *et al.*, 1989) which allows us to concentrate

on only those problems which are important from the perspective of constructing the STI-R.

Just as in the original version of the STI, in developing the STI-R we decided to limit the number of CNS properties to the ones proposed by Pavlov (1951–1952), i.e. to *strength of excitation*, *strength of inhibition*, and *mobility*. *Balance* of the CNS properties, regarded by Pavlov as the ratio between strength of excitation and strength of inhibition, is considered in our conceptualization as a secondary property.

The neo-Pavlovian typologists working under Teplov (Nebylitsyn, 1972; Nebylitsyn and Gray, 1972; Teplov, 1964) enlarged the number of CNS properties as compared with Pavlov's original theory of higher nervous activity. Such properties as lability, dynamism, activatability, and concentratability of the CNS have been distinguished. These properties as well as their relation to the basic Pavlovian CNS properties have been described in detail elsewhere (Mangan, 1982; Nebylitsyn, 1972; Strelau, 1983). The main reason for not including the neo-Pavlovian properties into the STI-R consists in the fact that experimental data collected by Teplov and his students (Nebylitsyn 1972; Nebylitsyn, Golubeva, Ravich-Shcherbo and Yermolayeva-Tomina, 1965; see also Strelau, 1983) do not allow us to conclude that these properties are orthogonal against the original ones. Moreover, the indicators of these properties are lacking in generalizability due to their low cross-situational consistency (Strelau, 1983, 1990a).

Our idea, also underlying the original version of the STI, was to develop an inventory which allows us to measure the CNS properties according to their original meaning as represented by Pavlov (1951–1952). However, the fulfilment of this goal is hardly possible, for at least two reasons. First, during the almost 30 years of studies Pavlov often changed his views regarding the understanding of the basic CNS properties and the ways of measuring them. Second, CNS properties were studied by Pavlov exclusively in dogs, mainly in laboratory settings based on the conditioned reflex paradigm. To avoid misunderstanding, the conceptualization of CNS properties underlying the construction of the STI-R is based mainly on one of Pavlov's last papers, *General Types of Higher Nervous Activity in Animals and Man*, published in 1935. This publication is considered by distinguished neo-Pavlovian typologists (Merlin, 1973; Nebylitsyn, 1972; Teplov, 1964) as the most systematic and full presentation of Pavlov's theory of the types of the central nervous system. Since the STI-R items refer to overt behaviour which has not much in common with conditioned reflexes (CR) as measured in Pavlov's laboratory, rather the ideas underlying the CR methods aimed at measuring a given CNS property and not the methods by themselves have been taken as a starting point for generating items.

It has to be stressed that Pavlov (1951–1952), when defining the basic CNS properties, did not refer to physiological mechanisms, as the names of these properties suggest. He characterized them from the *functional* point of view, stressing the role they play in the process of the individual's adaptation to the environment, thus taking in fact a position of a behaviourist in defining and studying the CNS properties (Strelau, 1983; Windholz, 1987), which he regarded, when referring to man, as temperament characteristics.

The CNS properties as understood by Pavlov (1951–1952) and operationalized in the STI-R, have been described in detail elsewhere (Strelau, 1983; Strelau *et al.*, 1989). Therefore, we will limit our description to the definitional aspects of the properties under discussion. In constructing the separate scales of the STI-R it was

assumed that the properties of the CNS, regarded by Pavlov (1951–1952) as general traits, reveal themselves in all kinds of behaviour, such as motor characteristics, verbal activity, emotional reactions, etc. (Strelau, 1983).

Strength of excitation refers, according to Pavlov, to the functional capacity of the CNS and manifests itself in the ability to endure intense or long-lasting stimulation without passing into protective (transmarginal) inhibition. It has been argued elsewhere (Strelau, 1983) that there exist different sources of stimulation, such as, for instance, situations, settings, tasks, as well as discrete stimuli characterized by a given degree of variation, novelty, intensity, complexity, and meaningfulness. The individual's own activity may also be regarded as a source of stimulation (Fiske and Maddi, 1961; Strelau, 1983). Different activities which carry varied aspects of threat, risk, and tension, and which have a direct impact in increasing the level of activation, are of special significance in generating stimulation.

In agreement with Pavlov (1951–1952) it is assumed that protective inhibition, used as the most spectacular measure of strength of excitation, reveals itself in the decrease or in the disappearance of reactions to strong or prolonged stimulation as well as in disturbances of behaviour (mostly emotional in character) that are a result of this stimulation.

Having the above described characteristic of strength of excitation in mind, we constructed the Strength of Excitation (SE) scale of the STI-R, taking into account the following seven definitional components of this property: (SE1) Threatening situations do not restrain high SE persons from a former planned activity (action). Furthermore, a high SE person is (SE2) Prone to undertake activity (actions) in highly stimulating conditions, and (SE3) Prefers to carry out risky and/or demanding activities. (SE4) Performance of activity under social and/or physical load does not evoke emotional disturbances in high SE persons. (SE5) In the case of activities or situations of high stimulative value, their efficiency of performance does not decrease essentially. (SE6) High SE individuals are resistant against fatigue when performing long-lasting and/or intensive activity. Finally, (SE7) They are able to react adequately under strong emotional tension.

It has to be noted that the number of facets (components) of strength of excitation has been enlarged in the STI-R as compared with the STI from four to seven. This is the result of taking into account a larger range of sources of stimulation as well as broader aspects of behavioural expressions of this CNS property.

Strength of inhibition refers, according to Pavlov's theory of types of the CNS published in the 1930s (Pavlov, 1951–1952), to conditioned inhibition, which develops during ontogenesis. It reveals itself in the ability to maintain a state of conditioned inhibition, such as extinction, differentiation, delay, and conditioned inhibition in its narrow meaning. The persistence of inhibition is one of the basic indicators of that property. This persistence is manifested in the amount of time the CNS is able to remain in the state of conditioned inhibition. According to Pavlov (1951–1952), the ease of evoking conditioned inhibition and the stability of conditioned inhibitory processes are also indicators of this CNS property.

In constructing the Strength of Inhibition (SI) scale, we did not refer directly to conditioned reflexes, but to behaviours and reactions in which the above-mentioned types of conditioned inhibition are assumed to be manifested. Thus, for example, we expected that individuals with a weak nervous system as regards inhibition would be unable to sustain conditioned inhibition, which results, among other things, in

the inability or difficulty to stop a given behaviour when needed or to change reactions (e.g. emotional expression) when required. In distinction from the SI scale of the original STI which tackled only three aspects of strength of inhibition (see Strelau *et al.*, 1989), the SI STI-R scale comprises five definitional components of this CNS property. High SI persons (SI1) Easily restrain from behaviours which, for social reasons, are not expected or not desired. (SI2) They do not have difficulty in waiting for a task performance when a delay in such performance is expected. (SI3) Once starting to solve a given task or to react to a given situation, they are able to interrupt the performance (reaction) when needed. (SI4) They are able to delay their reactions to acting stimuli if this is required by the circumstances, and (SI5) They are able to hold back their expression of emotions when required.

Mobility of nervous processes has been defined by Pavlov (1951–1952) as the ability of the CNS to respond adequately as soon as possible to continuous changes in the environment. It has to be distinguished from lability, the latter being a CNS property introduced by Teplov (1964) and characterized by the speed with which the processes of the CNS are generated and terminated. The ability to react quickly and adequately to changes in the surroundings was measured in Pavlov's laboratory mainly by means of the so-called alteration method (see Strelau, 1983). The essence of this method consists in measuring the speed of elaborating adequate conditioned reflexes to changes in the signal value of conditioned stimuli. Taking as a point of departure Pavlov's definition of mobility, many behaviours and situations may be generated in which this CNS property is manifested. An *ex post* analysis of the MO scale of the original STI has shown that the items of this scale refer to both—mobility and lability of the CNS. The MO scale of the STI-R is destitute of this disadvantage and refers to mobility only, comprising the following five definitional components. A highly mobile person: (MO1) Reacts adequately to unexpected changes in the environment; (MO2) Adapts quickly to new surroundings; (MO3) Passes easily from one activity to another; (MO4) Changes mood lightly from positive to negative and vice versa, according to the meaning of the situation; and (MO5) Prefers situations which require different activities to be performed simultaneously.

These 17 components (7—SE, 5—SI, and 5—MO) distinguished on the basis of Pavlov's definitions of the CNS properties constituted the basis for the generation of items to be included into the STI-R. For reasons mentioned before, no definitional components have been separated for the equilibrium (balance) of nervous processes. The measure of balance is limited to a purely statistical procedure, owing to the fact that this CNS property is the ratio between strength of excitation and strength of inhibition.

REASONS FOR DEVELOPING THE STI-R

Some reasons for constructing the STI-R have already been mentioned in the Introduction, thus only additional arguments will be given here.

The most convincing argument for the development of the STI-R came from our content analysis of the relations between the original STI items and the definitional components (facets) of the CNS properties. In the construction of the original STI items, global scale definitions were used as guidelines for item nomination. This strategy may favour some definitional components more than others, resulting in divergent sample sizes of items for some components.

In a first inspection of the original STI items, three of us, working as judges, tried to assign the items of a given STI scale to the components of the respective concept. Taking full agreement between the judges as the criterion, we were able to classify 22 from the 44 SE items to one of the seven SE components, and 39 items of the 44 SI items could be assigned unambiguously to the five SI components. As regards the MO scale, including altogether 46 items, 23 items were assigned to the five MO components.

The distribution of items representing the components turned out to be very uneven. For the seven SE components the number of items varied from 1 to 5, for the SI scale from 1 to 15, and for the MO scale from 3 to 6. These uneven distributions, especially for the SI scale, should be judged as unsatisfactory.

In the flood of psychometric tools aimed at measuring temperament traits, the question arises as to whether there is a need to enlarge the number of inventories aimed at assessing this domain of behaviour characteristics.

Most of the psychometric techniques used in temperament research are aimed at diagnosing the behaviour characteristics under discussion in children (see Hubert, Wachs, Peters-Martin and Gandour, 1982; Strelau, 1990b), whereas the STI-R refers to adolescents and adults. Among the existing questionnaires for adults, only the STI is aimed at measuring traits which refer to the Pavlovian concept of temperament, this being the most discriminant feature of our inventory. A few examples will illustrate this statement. Among the temperament dimensions extracted by Buss and Plomin (1984) in their EAS Temperament Survey, only activity refers indirectly to strength of the nervous system. This survey does not touch at the domain of strength of inhibition and mobility. The Affect Intensity Measure Inventory constructed by Larsen and Diener (1987) refers only to the intensity aspect of emotions whereas strength of excitation comprises the intensity characteristics of all types of behaviour. The Stimulus Screening Questionnaire developed by Mehrabian (1977) as a measure of temperament is aimed at diagnosing the following three traits—arousability, pleasure, and dominance. Among them, only arousability has much in common with the concept of strength of excitation, but none of the three temperament traits refers to the other Pavlovian CNS properties. The Guilford–Zimmerman Temperament Survey (Guilford, Zimmerman and Guilford, 1976) belongs to the most popular diagnostic tools aimed at measuring temperament. It comprises ten dimensions, some of which refer rather to the personality domain (e.g. objectivity) and hardly can be accepted as temperament characteristics [see for the distinction between personality and temperament Strelau (1987)].

Many more examples may be given to show that the STI-R has its specificity and cannot be replaced by other psychometric measures of adults' temperament described in the literature. Maybe the most important advantage of the STI-R is the fact that this inventory allows us more efficiency than has been possible up to now to search for links between research on temperament conducted in the West and studies in this area as represented in Eastern Europe.

EMPIRICAL STRATEGIES IN CONSTRUCTING THE STI-R

As mentioned before, the empirical strategies chosen for our study were guided by the suggestions made by Angleitner *et al.* (1986). Our approach may be classified

as a rational–theoretical construction strategy. We do not believe that a purely empirical selection of the STI items, by choosing the best ones according to item, and item–scale characteristics will automatically result in a more reliable and valid diagnostic instrument. In a blind empirical selection of items, some of the definitional components of the CNS properties may not be represented at all.

Generation of items by experts

Relying on the merits of the intuitive, rational scale construction as proposed by Jackson (1970), we started with the component analysis of the Pavlovian CNS properties concepts. The 17 components (facets) were used as a starting point for item writing. In contrast to the original STI, which was first constructed in Polish, the set of items for STI-R was formulated in German. Each of us generated at least five items for each component. For item writing it was agreed that some basic rules should be followed. The items should be: (1) short and clearly understandable; (2) free from extreme levels of social desirability; (3) diverse in content so as to cover the whole universe of human conduct; (4) applicable to adults in different cultures and not biased towards particular populations, for example, college students or males; (5) logically related to the construct under consideration and at the same time not converged with similar but irrelevant constructs; and (6) balanced in their keying. Some of these rules should ensure item samples showing a considerable degree of substantive validity as proposed by Loevinger (1957).

The process of item writing was done independently by each of us. Altogether 377 items were generated (152 for the SE scale, 113 for SI, and 112 for MO), including 15 items from the original STI.

Items to facets sorting by experts for the STI-R

The items which were nominated for the respective components of the separate STI-R scale definitions were scrutinized for each component regarding their logical item–component relationship. Only items for which full agreement among the four judges was reached were selected. Again, the six rules mentioned above were used as criteria. The results of this judgmental procedure are given in Table 1, which shows the distribution of items for the respective components passing the presented criteria. The new items for the SE scale exhibited a distribution between 10 and 16 items for the seven facets comprising altogether 90 items. In general, they are well balanced regarding their keying. However, it turned out that for some facets it seemed extremely difficult to write negatively keyed items. The pool contains 16–18 items for each SI component, giving a total of 84 items. This set of items is very well balanced. For the MO facets the groups of items varied from 11 to 19 items. The MO items are also satisfactorily balanced. In total, the new STI-R item pool contains 252 items, among which there are 129 items positively and 123 negatively keyed.

Our current STI-R item pool contains more items than intended for the final version of the inventory for at least three reasons: (1) each component of the scale definition should be represented; (2) a social desirability scale based on extreme social desirable STI-R items will be constructed; and (3) some items will be eliminated in the forthcoming steps of scale construction because of unsatisfactory empirical item characteristics.

The control of social desirability of the STI-R items

Item formulations differ in the degree in which they evoke social–culturally determined values. In general, subjects tend to select social desirable responses. It is therefore important to consider the social desirability values of the items in the construction process. Items with extreme social desirability can be used as the basis for a Social Desirability (SD) scale.

The aim of the next step in constructing the STI-R was: (1) to explore the distribution of social desirability values of the 252 STI-R items; and (2) to construct a SD scale.

Method and procedure

The STI-R items were written separately on cards. For social desirability judgement the instruction proposed by Edwards (1957) was adapted. However, in the instruction, importance was given to judge the items in terms of whether the subject considered them as desirable or undesirable for him/herself. The judgement was done on the basis of a sorting procedure, using a 9-point Likert scale (from 1 = 'Extremely undesirable' to 9 = 'Extremely desirable'). For this sorting, boxes were presented in front of the subject. They were labelled from left to right: extremely, strongly, moderately, and mildly undesirable; neutral; and mildly, moderately, strongly, and extremely desirable. In addition to the STI-R items, 31 items were selected from the German Social Desirability Response Set (GSDRS) scale, developed by Schmidt and Vorthmann (1971). This procedure was aimed to give some anchoring points for evaluating the social desirability saturation of the STI-R items. The whole set of items was randomized for each subject. The task required about one hour's work. Twenty subjects volunteered (ten men and ten women, aged from 25 to 63).

Results

As an indicator of rater agreement, the coefficient ICC [2.20] [according to the taxonomy by Shrout and Fleiss (1979)] for the reliability of 20 raters and, for further differentiation, ICC [2.1] as an estimate of the agreement of a single rater were computed. These values for the STI-R scale components, STI-R scales, the whole STI-R item set, and the GSDRS scale are given in Table 1. The means and standard deviations of the social desirability rating for the STI-R scales and facets and for the GSDRS scale are also given in Table 1.

A comparison of the 31 items of the GSDRS scale with the whole STI-R item set revealed that the rater agreement was slightly higher for the GSDRS scale (ICC [2.20] = 0.97, ICC [2.1] = 0.61) than for the STI-R item set (ICC [2.20] = 0.95, ICC [2.1] = 0.46). Furthermore, for evaluating the accuracy of our raters the point biserial correlation between the scoring directions of the GSDRS scale items with the mean social desirability ratings of this item set was calculated. The correlation coefficient reached a value of 0.93, indicating that our raters worked reliably and accurately.

The rater agreement within the STI-R total scales ranged between 0.93 and 0.95 (ICC [2.20]) and between 0.41 and 0.51 (ICC [2.1]). The ICC [2.20] of the 17 STI-R components ranged between 0.81 (SE2) and 0.98 (SE4, SE7, MO1). The ICC [2.1] of the 17 components, however, displayed substantial differences in the rater agreements. The lowest agreement for a single rater occurred in the SE2 subscale items

Table 1. The 252-item pool of the STI-R: item distribution, direction of keying, and social desirability statistics

Facet/ scale	No. of items	Key +	Key −	*M*	*s*	ICC [2.1]	ICC [2.20]
SE1	12	7	5	6.34*	2.04	0.38	0.92
SE2	13	7	6	5.75*	2.07	0.17	0.81
SE3	16	11	5	6.17*	2.07	0.30	0.89
SE4	10	4	6	7.33	1.61	0.69	0.98
SE5	14	6	8	7.03	1.72	0.59	0.97
SE6	11	5	6	6.64†	1.96	0.45	0.94
SE7	14	6	8	7.17	1.60	0.67	0.98
SE	90	46	44	6.63*	1.87	0.47	0.95
SI1	17	10	7	6.54†	1.96	0.42	0.94
SI2	16	9	7	6.08*	1.73	0.35	0.91
SI3	16	9	7	6.30*	1.71	0.43	0.94
SI4	18	9	9	6.46*	1.67	0.50	0.95
SI5	17	7	10	6.39*	1.89	0.41	0.93
SI	84	44	40	6.35*	1.79	0.41	0.93
MO1	18	10	8	7.11	1.44	0.68	0.98
MO2	19	9	10	6.74†	1.61	0.59	0.97
MO3	18	11	7	6.45*	1.76	0.35	0.91
MO4	12	4	8	6.30*	1.76	0.38	0.92
MO5	11	5	6	6.05*	1.84	0.30	0.89
MO	78	39	39	6.53*	1.68	0.51	0.95
Total pool	252	129	123	6.50*	1.78	0.46	0.95
GSDRS	31	16	15	7.22	1.64	0.61	0.97

Note: Significant differences between means of facets or scales and the mean (7.22) of the GSDRS (two-tailed *t*-tests): $*p < 0.001$; $\dagger p < 0.01$. *M* = mean; *s* = standard deviation; SE = strength of excitation; SI = strength of inhibition; MO = mobility; GSDRS = German Social Desirability Response Set; ICC = intra-class correlation coefficient.

(ICC [2.1] = 0.17), the highest in SE4 (ICC [2.1] = 0.69). In general, it can be concluded that rater agreement proved to be reliable.

In 226 of 252 cases (89.7 per cent), the characterization of the individual items as desired or undesired features, expressed by the mean judgement of the 20 raters, was in accordance with the item keying for the STI-R scales. In order to compare social desirability of the components, for further calculations, all undesired items (mean < 5) were recoded and the mean values for components, scales, the GSDRS, and the complete STI-R were computed (see Table 1).

With regard to sex, no significant differences were detected. A two-tailed *t*-test was carried out for differences between the SDR means of the facets or scales and the GSDRS. With means greater than 7, the components SE4, SE5, SE7, and MO1 did not differ significantly from GSDRS. However, the three scales SE, SI, and MO as well as the entire STI-R showed significantly ($p < 0.001$) lower means than GSDRS.

Now the item with the highest SD value was removed from each component and added to the SD scale introduced above. Out of these 17 items, five were again excluded and reassigned to their components in order to increase the internal

consistency of the SD scale as measured by Cronbach alpha. Cronbach alpha reached 0.69, the mean corrected item–total correlation was 0.33 (see Table 2). During the further procedure of item selection, only those STI-R items were to be excluded, with regard to social desirability, which correlated higher with SD than with their corresponding scale. This criterion should help to reduce the SD-saturated part of the variance.

Item and scale characteristics of the STI-R

Method and procedure

The STI-R was administered to a sample of 510 subjects (340 women, 170 men), recruited in Bielefeld, Düsseldorf (Germany) and Graz (Austria) in university and school courses and by announcements in local newspapers (mixed sample). The whole sample was heterogeneous with respect to profession and age. The mean age was 30.3 with a standard deviation of 13.4. The mean age of men was 32.4 ($s = 13.5$); the mean age of women 29.4 ($s = 13.3$). The age of this sample ranged from 14 to 81 years. For the 17 components plus SD (overall 240 STI-R and 12 SD items), reliability coefficients (Cronbach alpha), corrected item–total correlations, item statistics, as well as correlations of the individual items with the components and the four scales (SE, SI, MO, SD) were computed[1].

It should be kept in mind that the 252-item version of the STI-R was considered only as a starting point for the item selection to construct scales with fewer items and improved itemmetric and scale values. Therefore, criteria for item selection to construct an improved version of the new 252-item-pool inventory—the STI-R—were set up. These criteria are explained below.

Item selection and development of the STI-R

For the development of a revised and reduced version of the STI (STI-R) an item was excluded if at least one of the following three criteria were met:

(1) corrected item–total correlation < 0.15;
(2) item correlation with corresponding scale less than correlation with the other scales;
(3) item correlation with corresponding scale not significant.

Furthermore, if the item correlation with another component of the corresponding scale happened to be greater than the correlation coefficient with the assigned component, the item was assigned to that higher correlated component. For the remaining corrected item set, all computations were performed again and the above-mentioned criteria were all checked anew. This procedure was repeated until no more corrections had to be carried out. A flow diagram of the item selection procedure is shown in Figure 1.

Results of the item selection procedure

From the 252-item pilot form of the STI-R, 86 items were excluded. From this pool of excluded items 32 did not reach a sufficient corrected item–total correlation

[1] The latter detailed item statistics and furthermore all sex-specific data as well as data based on replication studies with the respective tables can be obtained from A. Angleitner.

value >0.15; 36 items were eliminated because of higher correlations with non-corresponding scales; two items were dropped for their non-significant correlation with their corresponding scales. By eliminating nine items with lowest (though sufficient) corrected item–total correlations from SI5, MO1, and MO2, the number of items per component was adjusted to a maximum of 12 without loss of reliability in any component. However, in components SE4 and MO5 the number of items fell below seven. Therefore, these components (totalling seven items) were eliminated completely for reasons of low reliability. The resulting total number of items in scales SE, SI, and MO is 53, 54, and 47, respectively. Together with the 12 SD items, this set of 166 items represents the STI-R.

The means and standard deviations of the components and scales of the STI-R are given in Table 2. In general, marked sex differences were found. An analysis of variance with the scales as dependent variables and sex as factor revealed significant sex differences for the SE and SI scales. Men scored higher than women. These differences were in most cases also replicated for the level of the components. For the SE components the only exception was SE2, where no significant sex differences were found. For the SI components in SI4 and SI5 significant differences and for the MO components a significant difference for MO3 turned up.

The item and scale characteristics of the STI-R are also shown in Table 2. The reliability coefficients of the SE, SI, and MO scales were comparable to the unreduced (252-item pool) STI-R scales, reaching values of 0.89, 0.85, and 0.89 for the respective scales. Whereas in the original STI over one-third of the items (including all scales) showed extreme endorsement frequencies above 75 per cent or below 25 per cent, the extreme endorsement frequencies for the STI-R items were as follows: SE = 16.7 per cent, SI = 13.2 per cent, and MO = 31.9 per cent. The percentages of corrected item–scale correlations below 0.20 for the STI-R were reduced to 7.5 per cent for SE, 13.0 per cent for SI, and 8.5 per cent for MO, compared with the original STI (28.6, 23.6, and 42.1 per cent). These item statistics document that the STI-R shows quite an improvement. Looking at the components of the STI-R, containing 7–12 items, it may be stated that with the exception of the components SE2, SE7, and MO3, showing a small decrease in the reliabilities, the reliability values of the remaining components increased as compared with the pilot form of the STI-R. In general, the mean item–total correlations also improved. The STI-R scales may also be considered as slightly more balanced in comparison with the pilot form of this inventory.

For the 166-item version (STI-R), again the social desirability rating values were calculated (see Table 3).

The correlations between the STI-R scales and the components are given in Table 4.

As Table 4 shows, there is a strong positive relationship of 0.56 between the scales SE and MO, whereas SE and SI (0.22), as well as SI and MO (0.30), show considerably lower coefficients. All three scales exhibit substantial correlations with social desirability. Therefore, one may hypothesize that the correlations between the STI-R scales are mediated by the social desirability saturation of the scales. By using the partial correlation technique, the effect of the social desirability variance was therefore partialed out. For the resulting correlational picture see the three values in parentheses in the upper triangle of Table 4. These coefficients demonstrate that there is also content variance involved in producing a partial correlation of 0.39 between the SE and MO scales.

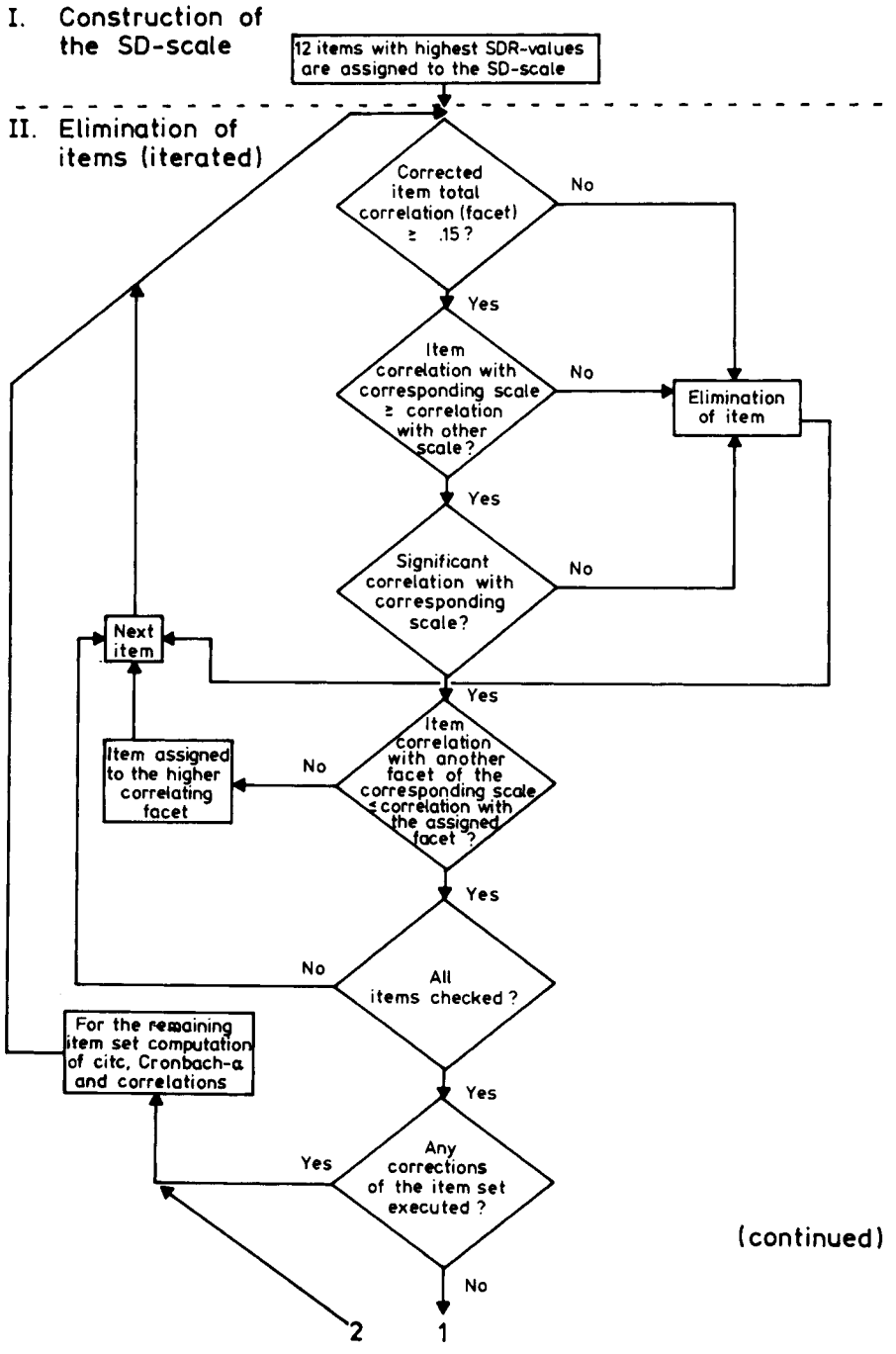
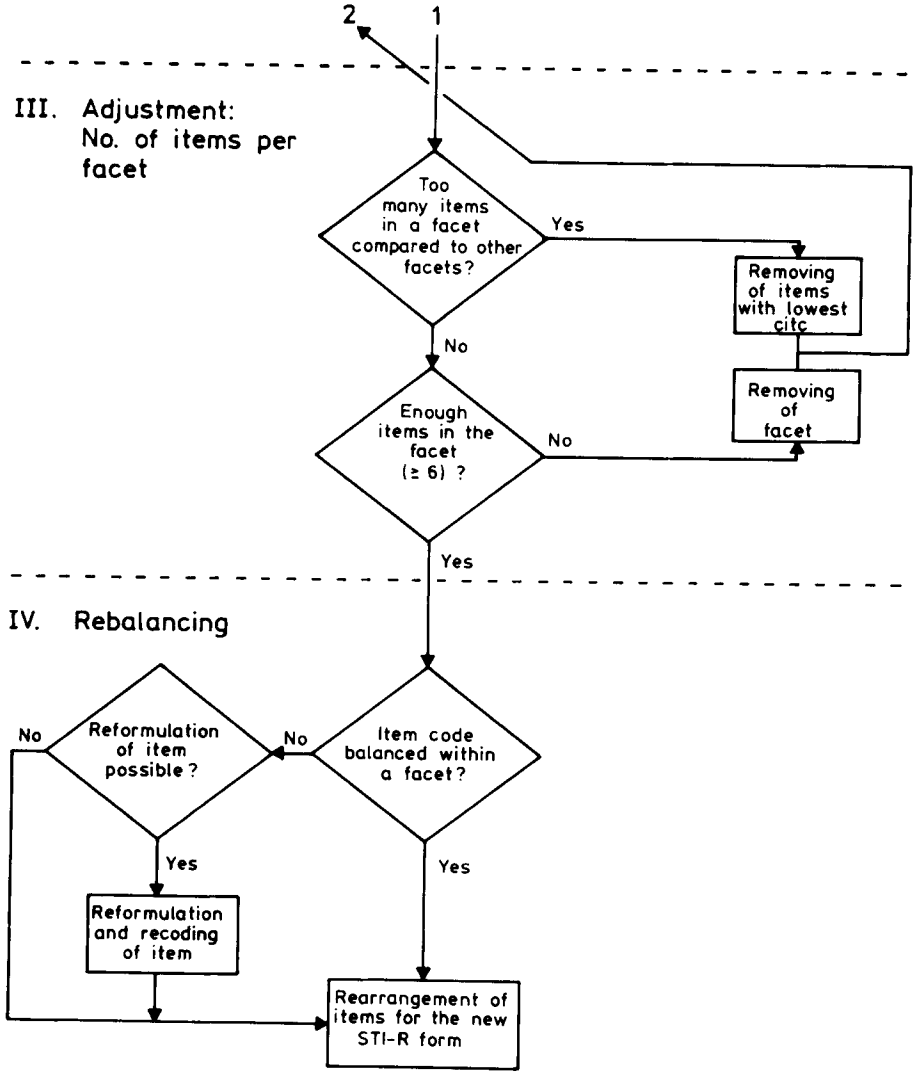


Figure 1. Flow diagram of the item selection procedure of the STI-R



Note.
citc = corrected item total correlation;
SDR = social desirability rating value of an item.

Figure 1 (*continued*)

Table 2. Psychometric characteristics of facets and scales of the STI-R (166 items)

Facet/ scale	*M*	*s*	No. of items	Cronbach α	Mean citc	Endorsement (%)
SE1	3.70	2.02	8	0.64	0.34	42
SE2	2.69	1.92	8	0.61	0.32	54
SE3	4.21	2.58	10	0.74	0.41	49
SE5	4.36	3.42	12	0.84	0.51	52
SE6	3.65	2.13	8	0.70	0.39	56
SE7	3.34	1.94	7	0.68	0.40	48
SE	21.92	9.56	53	0.89	—	50
SI1	8.06	2.42	12	0.65	0.30	42
SI2	4.09	2.49	9	0.75	0.45	48
SI3	5.15	2.26	9	0.69	0.37	49
SI4	6.34	2.61	12	0.66	0.30	48
SI5	7.21	2.96	12	0.79	0.43	38
SI	30.93	8.42	54	0.85	—	45
MO1	9.03	2.78	12	0.79	0.44	52
MO2	8.49	2.92	12	0.80	0.45	60
MO3	6.34	2.75	12	0.70	0.33	40
MO4	6.60	3.04	11	0.80	0.47	52
MO	30.48	8.56	47	0.89	—	51
SD	9.61	2.22	12	0.69	0.33	42

Note: citc = corrected item–total correlation.

Further inspection of Table 4 reveals that the strong correlation between SE and MO is also documented in the higher correlations between their respective components. However, the correlations of SE, SI, and MO with their corresponding components are without exception higher than with their non-corresponding scales and the SD scale. For detailed information about the inter-facet correlations see Table 5. Generally, the mono-facet correlations tend to show higher correlations than the respective values of the hetero-facet blocks. However, especially within the SE–MO block, there are some high values that document the above-mentioned high SE–MO scale correlation.

Development of a STI-R Short form (STI-RS)

For the purposes of research it seemed desirable to have a *short form* of the revised Strelau Temperament Inventory (STI-RS) at hand. The STI-RS was constructed in the following manner. The four best items of each component were selected according to their scale item–total correlations. This principle should guarantee that each component is reflected in the scale definitions. This procedure allows 24 items for the SE scale, 20 for the SI scale, and 16 for the MO scale to be separated. For reasons of having comparable lengths of scales, SI and MO were enlarged to 24 items by selecting the next best items of the remaining STI-R set.

As expected, the STI-RS corresponds closely to the STI-R. For the scales SE, SI, and MO, the obtained correlations are 0.95, 0.92, and 0.96. The Cronbach alpha values of the RS scales (see Table 6), varying between 0.80 and 0.88, may be judged

Table 3. Social desirability statistics for the STI-R

Facet/scale	No. of items	*M*	*s*
SE1	8	6.03*	2.15
SE2	8	5.63*	2.00
SE3	10	6.19*	2.09
SE5	12	6.80‡	1.76
SE6	8	6.47†	2.03
SE7	7	6.92‡	1.76
SE	53	6.36*	1.96
SI1	12	6.47†	2.02
SI2	9	5.88*	1.82
SI3	9	6.37*	1.67
SI4	12	6.55*	1.72
SI5	12	6.42†	1.94
SI	54	6.36*	1.84
MO1	12	7.08	1.47
MO2	12	6.90	1.65
MO3	12	6.55†	1.81
MO4	11	6.20*	1.79
MO	47	6.69*	1.68
SD	12	7.94*	1.09
Total pool	154* #	6.47* #	1.83 #
GSDRS	31	7.22	1.64

Note: Significant differences between means of facets or scales and the mean (7.22) of the GSDRS (two-tailed *t*-tests): $*p < 0.001$; $\dagger p < 0.01$; $\ddagger p < 0.05$; # SD scale not included.

as highly satisfactory, especially if one considers that with roughly half the number of items used in the original STI considerably higher reliability coefficients were achieved.

The means and standard deviations of the RS form are also given in Table 6. By applying an ANOVA with Sex as factor significant sex differences for all four STI-RS scales were revealed, documenting higher values for men compared with women. The percentages of items showing extreme endorsement frequencies were 16.7 per cent for SE, 16.7 per cent for SI, and 29.2 per cent for MO. There were almost no items which exhibited corrected item–scale correlations below 0.20 (SE: 4.2 per cent; SI: 0 per cent; MO: 0 per cent). Also the correlations between the scales (see Table 7) remain similar to those reported for the STI-R.

Replication studies

For replicational reasons two more samples were tested with the STI-R. The Düsseldorf sample contained 132 subjects (60 males; 72 females) ranging from 18 to 70 years, with a mean age of 30.8 years (for males 32.2; for females 29.7) and a standard deviation of 12.1 (for males 12.2; for females 11.9). Of this sample 27 per cent were students, 10 per cent were working in a health profession, 25 per cent were distributed over 18 different professions, and 38 per cent did not name their profession. The

Table 4. Intercorrelation coefficients (Pearson) of the STI-R scales and facets

Scales/facets	SE	SI	MO	SD
SE		(01)	(39)	48
SI	22		(04)	45
MO	56	30		61
SE1	*43*	00	30	27
SE2	*39*	−04	24	08
SE3	*49*	03	39	32
SE5	*57*	28	42	36
SE6	*55*	20	43	42
SE7	*56*	40	47	46
SI1	−09	*43*	10	23
SI2	22	*40*	30	29
SI3	19	*35*	28	31
SI4	17	*52*	09	29
SI5	23	*43*	23	34
MO1	46	28	*61*	58
MO2	35	09	*52*	44
MO3	51	28	*49*	48
MO4	36	26	*46*	35

Note: SD is partialed out for the three values in parentheses. The italicized values are part–whole corrected coefficients. All decimal points are omitted. $n = 506$; $r > 0.13$: $p < 0.001$.

Table 5. Intercorrelation coefficients (Pearson) of the STI-R facets

Facet	SE1	SE2	SE3	SE5	SE6	SE7	SI1	SI2	SI3	SI4	SI5	MO1	MO2	MO3
SE2	15													
SE3	49	23												
SE5	19	48	26											
SE6	31	19	33	47										
SE7	34	15	32	52	47									
SI1	−17	−14	−14	02	−03	05								
SI2	04	01	06	24	17	32	21							
SI3	04	01	02	26	14	26	23	30						
SI4	02	−01	−00	20	17	28	42	28	26					
SI5	06	−02	13	18	19	36	25	29	19	39				
MO1	27	16	34	33	35	40	13	31	17	12	19			
MO2	24	16	36	18	22	28	01	16	08	−04	09	53		
MO3	24	20	27	43	48	39	06	21	38	14	15	46	35	
MO4	14	22	19	31	26	34	09	23	21	06	25	40	35	36

Note: $n = 506$; $r > 0.14$: $p < 0.001$. All decimal points are omitted.

Bielefeld-2 sample was recruited again by announcements in local newspapers. This sample was heterogeneous with respect to profession. 19.7 per cent were students, 78.7 per cent were distributed over 25 different professions, and 1.6 per cent did not name their profession. This sample consisted of 122 subjects (56 males; 66 females) ranging from 15 to 81 years, with a mean age of 32.0 years (33.9 for males; 30.4

Table 6. STI-RS (84 items): psychometric characteristics

Scale	No. of items	*M*	*s*	Cronbach α	Mean citc	Endorsement (%)
SE	24	9.91	5.28	0.84	0.39	47
SI	24	12.66	4.78	0.80	0.35	42
MO	24	16.06	5.67	0.88	0.46	45
SD	12	9.61	2.22	0.69	0.33	42

Note: $n = 500$; citc = corrected item–total correlation.

Table 7. Intercorrelation coefficients (Pearson) of the STI-RS scales

Scale	SE	SI	MO
SE		0.12	0.36
SI	0.30		0.12
MO	0.55	0.34	
SD	0.50	0.42	0.61

Note: SD is partialed out in the upper triangle; $n = 506$; $r > 0.13$: $p < 0.001$.

for females) and a standard deviation in age of 13.8 years (15.3 for males; 12.2 for females).

Results of the STI-R

The reliability coefficients of the STI-R scales for the two samples were quite similar, ranging from 0.83 to 0.90 for the Bielefeld-2 sample and from 0.86 to 0.87 for the Düsseldorf sample. For the SD scale the respective alpha values were 0.70 and 0.66. In Table 8 detailed information concerning the reliabilities, mean corrected item–scale correlations, mean endorsement frequencies as well as scale means and standard deviations is given. For the replication samples a multivariate analysis of variance of the scales and components by the factors Sample and Sex were computed. For the STI-R–SE scale highly significant ($p < 0.01$) sex differences were found, showing men scoring higher than women. These differences hold for the SE components too, where men exhibited higher values also. For the SI components, only SI5 showed a marked sex difference in the above-mentioned direction.

For the facets again a higher variation for the reliability coefficients was found, ranging from 0.60 to 0.87 for SE in Bielefeld-2 and from 0.56 to 0.83 for the SE scale in the Düsseldorf sample. The respective values for the SI scale were 0.58–0.83 (Bielefeld-2) and 0.53–0.82 (Düsseldorf); for the MO scale these values varied between 0.65 and 0.82 (Bielefeld-2), and 0.65 and 0.80 (Düsseldorf).

The correlations between the scales indicated again that the SE and MO scales are correlated. SE and MO showed a correlation of 0.61 in the Bielefeld-2 sample. In the Düsseldorf sample, the correlation between these scales reached 0.48. However, whereas SE and SI were quite similar in their correlations in the replication samples (0.17 and 0.11, Bielefeld and Düsseldorf samples, respectively), the correlations between SI and MO differed in the two samples (0.13 Bielefeld-2 and 0.43

Table 8. Psychometric characteristics of facets and scales of the STI-R (166 items): replication studies

Facet/ scale	No. of items	Bielefeld-2 sample (n = 116)					Düsseldorf sample (n = 132)				
		M	s	Cronbach α	Mean citc	Endorse-ment (%)	M	s	Cronbach α	Mean citc	Endorse-ment
SE1	8	4.41	1.98	0.60	0.32	44	3.68	1.95	0.60	0.31	45
SE2	8	2.09	1.93	0.67	0.38	57	2.48	1.93	0.64	0.34	57
SE3	10	4.84	2.86	0.79	0.48	51	4.58	2.65	0.74	0.40	49
SE5	12	4.75	3.63	0.87	0.57	56	4.47	3.33	0.83	0.49	53
SE6	8	4.12	2.06	0.67	0.37	59	3.64	2.06	0.68	0.37	57
SE7	7	3.68	1.95	0.69	0.40	51	3.19	1.74	0.56	0.30	47
SE	53	24.00	9.91	0.90	—	53	21.98	9.08	0.87	—	51
SI1	12	8.24	2.17	0.58	0.25	45	3.82	2.34	0.63	0.27	44
SI2	9	4.46	2.50	0.75	0.45	49	4.44	2.62	0.78	0.50	50
SI3	9	4.96	2.07	0.63	0.32	51	5.25	2.23	0.67	0.35	48
SI4	12	6.53	2.52	0.63	0.28	51	6.32	2.28	0.53	0.21	52
SI5	12	7.40	3.22	0.83	0.50	41	6.67	3.12	0.82	0.48	42
SI	54	31.45	7.79	0.83	—	47	31.62	8.61	0.86	—	47
MO1	12	9.32	2.86	0.82	0.50	54	9.51	2.65	0.80	0.46	53
MO2	12	8.62	2.94	0.81	0.47	59	8.70	2.84	0.79	0.45	61
MO3	12	6.93	2.50	0.65	0.30	37	6.52	2.58	0.65	0.30	39
MO4	11	6.89	2.98	0.81	0.48	53	6.61	2.80	0.77	0.43	51
MO	47	31.69	8.64	0.90	—	51	31.28	7.92	0.87	—	51
SD	12	10.03	2.11	0.70	0.35	43	9.87	2.06	0.66	0.31	41

Note: citc = corrected item–total correlation.

Düsseldorf). Furthermore, the scales as well as almost all components showed positive correlations with the SD scale. But compared with the construction sample, all components correlated higher with their own scale than with the non-corresponding scales and the SD scale. No marked sex differences were found concerning facet and scale correlations within each sample. The inter-facet correlations of the STI-R for the two replication samples were comparable to the coefficients obtained in the construction sample (as shown in Table 5). A rank correlation (Spearman) between the construction and the Bielefeld-2 sample displayed a coefficient of 0.87. The correlation between the former and the Düsseldorf sample was 0.80. For the correlation between the two replication samples a value of 0.70 ($n = 105; p < 0.001$) was obtained.

Results of the STI-RS

The reliability values for the replication samples are documented in Table 9. The Cronbach alpha coefficients were again highly similar for the two samples. They ranged from 0.79 to 0.89 in the Bielefeld-2 sample and from 0.79 to 0.88 in the Düsseldorf sample. Almost no differences were found considering the mean corrected item–scale correlations, the mean endorsement frequencies, as well as the scale means and standard deviations.

The correlations between the STI-RS scales in the replicational studies were as follows: in both samples (first values represent Bielefeld-2), SE and MO showed a high correlation (0.60 and 0.50); the correlations between SI and MO differed somewhat in the two samples (0.19 and 0.40); for SE and SI, the correlations were quite similar (0.18 and 0.16). As expected, the STI-RS scales also showed significant positive correlations with the SD scale.

The STI-RS scales correlated highly with the STI-R scales. The values were 0.92 (SI) and 0.96 (SE and MO) for the Bielefeld-2 sample, and 0.93 (SI) and 0.96 (SE and MO) for the Düsseldorf sample. The STI-RS scales showed higher correlations with corresponding STI-R facets than with non-corresponding ones.

A multivariate analysis of variance was computed with the STI-RS scales as variables and using the Replication samples and Sex as factors. This analysis confirmed the already established sex differences for the SE scale documenting higher values for men compared with women.

THE FINAL VERSIONS OF THE STI-R AND STI-RS

For the final version of the STI-R the set of 166 items was newly balanced. Two items for SE, two for SI, and three for MO were reformulated; however, without changing the items' meaning. Furthermore, the answering format was changed to a 4-point Likert scale, because some subjects reported difficulty in answering 'Yes' or 'No', especially to negative item formulations. The labels for the new answering format were: 'Fully agree', 'Agree', 'Disagree', and 'Disagree completely'. Finally, the resulting items of the four scales—SE, SI, MO, and SD—were rearranged in an alternating succession.

In a retest study, the correspondence between this final STI-R version and the 252-item pilot form of the STI-R was examined. In a further study, the stability of the STI-RS was analysed by giving the subjects the STI-RS twice.

Table 9. Psychometric characteristics of facets and scales of the STI-RS (84 items): replication studies

		Bielefeld-2 sample (n = 116)					Düsseldorf sample (n = 132)				
Scale	No. of items	M	s	Cronbach α	Mean citc	Endorse-ment (%)	M	s	Cronbach α	Mean citc	Endorse-ment
SE	24	10.98	5.12	0.83	0.38	50	10.10	4.76	0.79	0.33	47
SI	24	13.06	4.63	0.79	0.34	44	13.42	4.78	0.80	0.35	43
MO	24	16.88	5.63	0.89	0.48	45	16.34	5.55	0.88	0.46	44
SD	12	10.04	2.11	0.70	0.35	43	9.87	2.06	0.66	0.31	41

Note: citc = corrected item–total correlation.

Table 10. Means, standard deviations, and reliabilities for STI-R and STI-RS facets and scales with 4-point answer format

Facet/scale	*M*	*s*	Cronbach *α*
SE1	18.77	4.22	0.77
SE2	15.32	4.77	0.78
SE3	21.91	5.08	0.73
SE5	25.75	6.09	0.86
SE6	19.31	4.36	0.77
SE7	16.88	3.43	0.66
SE	118.43	19.74	0.91
SI1	33.92	5.62	0.78
SI2	21.13	4.34	0.75
SI3	25.22	4.36	0.72
SI4	31.41	5.32	0.76
SI5	31.91	6.80	0.88
SI	143.64	19.35	0.91
MO1	34.94	5.44	0.78
MO2	34.57	6.26	0.82
MO3	31.75	5.35	0.69
MO4	29.23	5.86	0.83
MO	129.75	18.66	0.91
SD	36.38	4.82	0.77
SE(RS)	52.64	10.53	0.88
SI(RS)	61.15	9.70	0.85
MO(RS)	66.88	11.71	0.91

Note: $n = 76$.

Method and procedure

Seventy-six subjects (21 men and 55 women) who had already been tested with the STI-R (252-item pool) 12 months before again filled out the final STI-R version. The sample was heterogeneous with respect to education and profession. Subjects were not paid for their participation and worked at home. Materials and detailed instructions were sent by mail. In the stability study of the STI-RS, 74 subjects (31 men and 43 women; aged from 17 to 68) participated. The time interval between the two test sessions was 4–6 weeks. In general, the subjects filled out the questionnaire in the presence of the experimenter.

Results

Reliabilities (Cronbach alpha) were computed for STI-R scales and facets. These coefficients, 0.91 (SE), 0.91 (SI), and 0.91 (MO), and, furthermore, the coefficients for facets (see Table 10) are higher than the respective values obtained in earlier studies.

The facet–scale correlations (see Table 11) for the 4-point Likert scale STI-R reveal greater discrimination compared with the Yes–No STI-R form administered

Table 11. Scale–facet correlations for STI-R with 4-point answer format

Facet/scale	SE	SI	MO	SD
SE				62
SI	22			57
MO	64	39		63
SE1	*58*	08	53	50
SE2	*46*	−16	22	10
SE3	*46*	−07	38	37
SE5	*70*	39	53	52
SE6	*59*	26	46	55
SE7	*65*	51	69	72
SI1	−12	*60*	12	31
SI2	23	*52*	40	41
SI3	10	*47*	39	37
SI4	24	*71*	25	54
SI	32	*36*	28	41
MO1	*59*	38	*74*	66
MO2	48	07	*58*	42
MO3	52	43	*64*	60
MO4	48	40	*63*	40

Note: All decimal points are omitted. $n = 76$; $r \geq 0.27$: $p < 0.01$; $r \geq 0.36$: $p < 0.001$. The italicized values are part–whole corrected coefficients.

12 months earlier, especially for the ratio between SE and MO. The convergence (Pearson correlations) between the two STI-R versions can be seen from Table 12. The correlations for the scales are 0.83 (SE), 0.68 (SI), 0.79 (MO), and 0.62 for the SD scale. The lowest convergence for the SE facets is 0.56 (SE1), the highest value 0.81 (SE3). For the SI facets these values range between 0.57 (SI3) and 0.68 (SI5), and for the MO facets between 0.57 (MO3) and 0.75 (MO2).

The inter–facet correlations of this STI-R version are given in Table 13. In most of the cases, the empirical relations between the facets as shown in Table 5 are replicated. The mono-facet correlations are higher than the hetero-facet correlations. However, some of the correlations between the components of SE and MO indicate again the lack of discrimination between SE and MO.

Results for the 4-point Likert scale STI-RS form were computed from STI-R items. Cronbach alpha coefficients for the STI-RS scales are 0.88 for SE, 0.85 for SI, and 0.91 for MO (see Table 10). The correlations between the STI-R and STI-RS scales with the 4-point answer format are 0.96 (SE), 0.96 (SI), and 0.97 (MO). Pearson correlations for the three STI-RS scales (with different answer formats) between the two test times are 0.83 (SE), 0.60 (SI), and 0.76 (MO). The stability scores increased when a 4-point answer format and a shorter time interval were used. The respective values are as follows: 0.88 (SE), 0.86 (SI), 0.88 (MO), and 0.84 (SD).

Generally, the results show high reliabilities and high facet–scale discriminance. If one considers the retest interval of about 12 months together with the changed answer format and some reformulations of items, as well as the stability values of the STI-RS scales, the retest 'convergence' is encouraging.

Table 12. Convergence between the STI-R with 2-point and 4-point answer format and 12-month retest interval

Scale/facet	Pearson correlations
SE	0.83
SI	0.68
MO	0.79
SD	0.62
SE1	0.56
SE2	0.70
SE3	0.81
SE5	0.68
SE6	0.63
SE7	0.63
SI11	0.62
SI2	0.61
SI3	0.57
SI4	0.62
SI5	0.68
MO1	0.73
MO2	0.75
MO3	0.57
MO4	0.66

Note: $n = 71$; $r \geq 0.28$: $p < 0.01$; $r \geq 0.36$: $p < 0.001$.

Table 13. Intercorrelation coefficients (Pearson) of STI-R facets (STI-R form with 4-point answer format)

Facet	SE1	SE2	SE3	SE5	SE6	SE7	SI1	SI2	SI3	SI4	SI5	MO1	MO2	MO3
SE2	31													
SE3	60	30												
SE5	41	51	22											
SE6	38	30	30	62										
SE7	45	30	32	72	57									
SI1	−18	−29	−26	07	−00	17								
SI2	07	−11	−16	45	33	43	37							
SI3	−02	−19	−15	26	29	24	54	50						
SI4	10	−14	07	32	30	42	59	47	36					
SI5	26	08	15	34	08	51	28	24	07	56				
MO1	44	15	36	52	49	65	11	47	38	27	20			
MO2	49	24	47	31	19	39	−09	14	10	06	07	62		
MO3	37	10	21	46	53	68	28	34	44	30	24	61	41	
MO4	39	21	19	44	33	55	11	37	36	20	40	56	45	58

Note: $n = 71$; $r > 0.23$: $p < 0.05$; $r > 0.28$: $p < 0.01$; $r > 0.36$: $p < 0.001$. All decimal points are omitted.

DISCUSSION

The results of the series of studies we conducted with the STI-R and STI-RS allow us to conclude that these inventories, as compared with the original STI, have better psychometric characteristics. The item statistics and reliability scores of the STI-R/STI-RS fulfil the criteria underlying the construction of personality inventories. The fact that the SE and MO scales correlate better with each other than expected from the factor analytic approach is, however, consistent with data referring to the original STI (see Strelau, 1983; Strelau *et al.*, 1989) as well as with the rational–theoretical strategy applied in our research. As mentioned above, it follows from Pavlov's theoretical considerations, supported by empirical studies, that these two properties correlate positively with each other. Since the STI-R and STI-RS are aimed at measuring the CNS properties within the Pavlovian tradition, there is no reason for undertaking trials which allow us to construct SE and MO scales which will be orthogonal to each other.

Since subjects reported some difficulty in answering the 'Yes' and 'No' format of the STI-R and STI-RS, it was decided to change the format of the inventories to a 4-point rating scale. Because of the increase in the reliability scores and a higher differentiation effect between components belonging to different scales when the 4-point Likert scale is applied, we recommend the use of the STI-R and STI-RS questionnaires with a 4-point rating scale format.

When detailed information is required in order to explain the nature and/or structure of the separate CNS properties, the full (166-item format) STI-R is proposed, because it allows the scores for the separate definitional components of the CNS properties to be measured.

If one compares the psychometric characteristics of the STI-R with the values of some other current temperament inventories, like the DOTS-R (Windle and Lerner, 1986) and EAS (Buss and Plomin, 1984), the STI-R will have its place showing comparable reliability coefficients. However, our cumbersome approach to construct a temperament inventory which should be relatively free from social desirability was not so successful. First, our SD scale does not show high reliability and, second, even if each item correlates more strongly with its own scale than with the SD one, the SD scale still shows considerable correlations with the STI-R content scales. It seems to be true that in temperament inventories scales for controlling response tendencies are seldom to be found. A more preferable strategy to the one we have used would be the application of an already established Lie or Infrequency scale (e.g. from the EPQ or PRF) together with the 252-item pool of the STI-R. The correlations between item responses with such a control scale might be considered for building a new STI-R Control scale. However, future research has to show whether such a strategy will yield to the construction of a successful Control scale.

In this paper we have concentrated on the construction and reliability issues. In a forthcoming report the validity of the STI-R and STI-RS will be shown (Ruch, Angleitner and Strelau, 1990).

ACKNOWLEDGEMENTS

We would like to thank William J. Corulla, Guus L. Van Heck, and the anonymous reviewers for their helpful comments on an earlier version of this manuscript.

The preparation of this article was supported in part by a grant from the University of Bielefeld (OZ 2794) to Alois Angleitner and by the Minister of National Education (RPBP III.25) to Jan Strelau.

The research reported in this paper is based on the German STI-R 252-item pool. The recommended German 166-item STI-R and the 84-item STI-RS inventories may be obtained by writing to Alois Angleitner or to Jan Strelau. The STI-R 252-item pool is considered as a starting point for developing the STI-R/STI-RS in different countries. The following translations of the 252-item pool are available: English, Italian, Korean, Polish, Russian, and Spanish. By writing to Alois Angleitner or to Jan Strelau, the Program for Cross-Cultural Approach and Co-operation in Constructing the STI-R together with the language versions may be obtained.

REFERENCES

Angleitner, A., John, O. P. and Löhr, F. J. (1986). 'It's what you ask and how you ask it: an itemmetric analysis of personality questionnaires'. In: Angleitner, A. and Wiggins, J. S. (Eds), *Personality Assessment via Questionnaires*, pp. 61–107, Springer-Verlag, New York.

Barnes, G. E. (1976). 'Individual differences in perceptual reactance: a review of the stimulus intensity modulation individual difference dimension', *Canadian Psychology Review*, **17:** 29–52.

Buchsbaum, M. S. (1978). 'Neurophysiological studies of reduction and augmentation'. In: Petrie, A. (Ed), *Individuality in Pain and Suffering*, 2nd ed., pp. 141–157, Chicago University Press, Chicago, IL.

Buss, A. H. and Plomin, R. (1984). *Temperament: Early Developing Personality Traits*, Erlbaum, Hillsdale, NJ.

Carlier, M. (1985). 'Factor analysis of Strelau's Questionnaire and an attempt to validate some of the factors'. In: Strelau, J., Farley, F. H. and Gale, A. (Eds), *The Biological Bases of Personality and Behavior: Theories, Measurement Techniques, and Development*, Vol. 1, pp. 145–160, Hemisphere, Washington, DC.

Claridge, G. (1985). *Origins of Mental Illness: Temperament, Deviance and Disorder*, Basil Blackwell, Oxford.

Edwards, A. E. (1957). *The Social Desirability Variable in Personality Assessment and Research*, Dryden Press, New York.

Eysenck, H. J. (1972). 'Human typology, higher nervous activity, and factor analysis'. In: Nebylitsyn, V. D. and Gray, J. A. (Eds), *Biological Bases of Individual Behavior*, pp. 165–181, Academic Press, New York.

Fiske, D. W. and Maddi, S. R. (Eds) (1961). *Functions of Varied Experience*, Dorsey Press, Homewood, IL.

Goldman, D., Kohn, P. M. and Hunt, R. W. (1983). 'Sensation seeking, augmenting–reducing, and absolute auditory threshold: a strength-of-the-nervous-system perspective', *Journal of Personality and Social Psychology*, **45:** 405–411.

Guilford, J. S., Zimmerman, W. S. and Guilford, J. P. (1976). *The Guilford–Zimmerman Temperament Survey Handbook: Twenty-five Years of Research and Application*, Edits Publishers, San Diego, CA.

Hubert, N. C., Wachs, T. D., Peters-Martin, P. and Gandour, M. J. (1982). 'The study of early temperament: measurement and conceptual issues', *Child Development*, **53:** 571–600.

Jackson, D. N. (1970). 'A sequential system for personality scale development'. In: Spielberger, C. D. (Ed), *Current Topics in Clinical and Community Psychology*, Vol. 2, pp. 61–96, Academic Press, New York.

Kohn, P. M., Cowles, M. P. and Lafreniere, K. (1987). 'Relationships between psychometric and experimental measures of arousability', *Personality and Individual Differences*, **8:** 225–231.

Larsen, R. J. and Diener, E. (1987). 'Affect intensity as an individual difference characteristic: a review', *Journal of Research in Personality*, **21:** 1–39.

Loevinger, J. (1957). 'Objective tests as instruments of psychological theory', *Psychological Reports*, **3:** 635–694.

Loo, R. (1979). 'Neo-Pavlovian properties of higher nervous activity and Eysenck's personality dimensions', *International Journal of Psychology*, **14:** 265–274.

Mangan, G. L. (1982). *The Biology of Human Conduct: East–West Models of Temperament and Personality*, Pergamon Press, Oxford.

Mehrabian, A. (1977). 'A questionnaire measure of individual differences in stimulus screening and associated differences in arousability', *Environmental Psychology and Nonverbal Behavior*, **1:** 89–103.

Merlin, V. S. (Ed) (1973). *Outline of the Theory of Temperament*, 2nd ed., Permskoye Knizhnoye Izdatelstvo, Perm (in Russian).

Nebylitsyn, V. D. (1972). *Fundamental Properties of the Human Nervous System*, Plenum Press, New York.

Nebylitsyn, V. D. and Gray, J. A. (Eds) (1972). *Biological Bases of Individual Behavior*, Academic Press, New York.

Nebylitsyn, V. D., Golubeva, E. A., Ravich-Shcherbo, I. V. and Yermolayeva-Tomina, L. B. (1965). 'A comparative study of short methods of measuring basic properties of the nervous system in man'. In: Teplov, B. M. (Ed.), *Typological Features of Higher Nervous Activity in Man*, Vol. 4, pp. 60–83, Prosveshcheniye, Moscow (in Russian).

Pavlov, I. P. (1951–1952). *Complete Works*, 2nd ed., SSSR Academy of Sciences, Moscow and Leningrad (in Russian).

Ruch, W., Angleitner, A. and Strelau, J. (1990). 'The Strelau Temperament Inventory-Revised (STI-R): validity studies'. Submitted for publication.

Schmidt, H. D. and Vorthmann, H. R. (1971). 'Eine Skala zur Messung der "sozialen Erwünschtheit" ' [Social desirability response set], *Diagnostica*, **17:** 87–90.

Shrout, P. E. and Fleiss, J. L. (1979). 'Intraclass correlations: uses in assessing rater reliability', *Psychological Bulletin*, **86:** 420–428.

Stelmack, R. M., Kruidenier, B. G. and Anthony, S. B. (1985). 'A factor analysis of the Eysenck Personality Questionnaire and the Strelau Temperament Inventory', *Personality and Individual Differences*, **6:** 657–659.

Strelau, J. (1969). *Temperament i Typ Ukladu Nerwowego* [Temperament and Type of Nervous System], Panstwowe Wydawnictwo Naukowe, Warsaw.

Strelau, J. (1972). 'A diagnosis of temperament by nonexperimental techniques', *Polish Psychological Bulletin*, **3:** 97–105.

Strelau, J. (1983). *Temperament—Personality—Activity*, Academic Press, London.

Strelau, J. (1987). 'The concept of temperament in personality research', *European Journal of Personality*, **1:** 107–117.

Strelau, J. (1990a). 'Are psychophysiological scores good candidates for diagnosing temperament/personality traits and for a demonstration of the construct validity of psychometrically measured traits?' Manuscript submitted for publication.

Strelau, J. (1990b). 'Renaissance in research on temperament: where to?' Manuscript submitted for publication.

Strelau, J. and Eysenck, H. J. (Eds) (1987). *Personality Dimensions and Arousal*, Plenum Press, New York.

Strelau, J., Angleitner, A. and Ruch, W. (1989). 'Strelau Temperament Inventory (STI): general review and studies based on German samples'. In: Spielberger, C. D. and Butcher, J. N. (Eds), *Advances in Personality Assessment*, Vol. 8, pp. 187–241, Erlbaum, Hillsdale, NJ.

Teplov, B. M. (1964). 'Problems in the study of general types of higher nervous activity in man and animals'. In: Gray, J. A. (Ed), *Pavlov's Typology*, pp. 3–153, Pergamon Press, Oxford.

Windholz, G. (1987). 'Pavlov as a psychologist: a reappraisal', *The Pavlovian Journal of Biological Science*, **22:** 103–112.

Windle, M. and Lerner, R. M. (1986). 'Reassessing the dimensions of temperamental individuality across the life span: the Revised Dimensions of Temperament Survey (DOTS-R)', *Journal of Adolescent Research*, **1:** 213–230.

Zuckerman, M. (1979). *Sensation Seeking: Beyond the Optimal Level of Arousal*, Erlbaum, Hillsdale, NJ.

Zuckerman, M., Kuhlman, D. M. and Camac, C. (1988). 'What lies beyond E and N? Factor analyses of scales believed to measure basic dimensions of personality', *Journal of Personality and Social Psychology*, **54:** 96–107.

RÉSUMÉ

Compte-rendu du développement de la révision du *Questionnaire Strelau Temperament* (STI-R). Le STI-R fournit une mesure des caractéristiques de base du système nerveux central (CNS): Force de l'excitation, force de l'inhibition et mobilité du CNS comme l'entend Pavlov. C'est sur la base d'une série de recherches que le développement des versions définitives du STI revu a connu différents stades. Les version suivantes sont examinées en detail: (1) une serie de départ de 252 items; (2) un STI-R de 155 items avec comme possibilités de réponse 'oui' et 'non'; (3) une forme courte (84 items) du STI-R (STI-RS) avec 'oui' et 'non' comme possibilités de réponse; (4) un STI-R comportant 155 items avec une échelle de Likert en quatre points; et (5) un STI-RS de 84 items avec une échelle de jugement en quatre points. Les caractéristiques psychométriques des versions successives du STI revu étaient améliorées à chaque étape. Celles-ci peuvent être considérées, en général, comme satisfaisantes. Les auteurs conseillent spécialement les versions nommées en (4) et (5). Ces dernières ont, entre autres, les scores de fiabilité les plus élevés. Elles sont considerées comme les formes définitives du STI-R et du STI-RS.

ZUSAMMENFASSUNG

Die Konstruktion eines revidierten Strelau Temperament Inventars (STI-R) wird berichtet. Gemäss dem Verständnis von Pavlov dient der STI-R der Messung dreier grundlegender Eigenschaften des zentralen Nervensystems (ZNS): Stärke der Erregung, Stärke der Hemmung und Mobilität des ZNS. Die folgenden STI-R Formen wurden auf Grund von mehreren Studien entwickelt: (1) eine 252-Item Ausgangsform des STI-R, (2) eine 166-Item STI-R Form mit 'ja' & 'nein' Antwort-Form, (3) eine Kurzform mit 84 Items mit 'ja' & 'nein' Antwort-Form (STI-RS), (4) eine 166-Item STI-R Form mit 4-stufiger Antwort-Form und (5) eine 84-Item Kurzform mit 4-stufiger Antwort-Form (STI-RS). Die psychometrischen Merkmale der verschiedenen Versionen des revidierten STI verbesserten sich Schritt für Schritt. Im allgemeinen können diese Merkmale als befriedigend beurteilt werden. Die Versionen (4) und (5) werden besonders empfohlen, da sie, unter anderem, die höchsten Reliabilitätswerte besitzen. Diese Testformen werden als entgültige Formen des STI-R und STI-RS betrachtet.