The Impact of Modified Instructions on Ego-Level Scores: A Psychometric Hazard or Indication of Optimal Ego Level?

Martine J. Drewes and P. Michiel Westenberg

Developmental and Educational Psychology,
Faculty of Social and Behavioral Sciences
Leiden University

Various modified instructions have different effects on the Washington University Sentence Completion Test for Ego Development (Loevinger, 1985). These effects were reviewed and 2 alternative explanations were explored: The “measurement unreliability explanation” versus the “optimal level explanation.” Both explanations were systematically studied in 2 test–retest experiments with the Sentence Completion Test for Children and Youths (Westenberg, Treffers, & Drewes, 1998). The modified instructions were to make a favorable impression on the tester (“fake good” condition) or to complete the sentences in as adult and mature a manner as one can (“be mature” condition). Both experiments were conducted with 9- to 15-year-old children and adolescents (N = 127, 128). As was anticipated, neither the fake good nor the be mature condition yielded higher (or lower) reliability indexes as compared to the standard instructions, hence discounting the measurement unreliability explanation. Also as expected, the fake good condition did not yield significantly higher ego-level scores, whereas the be mature condition did yield significantly higher ego-level scores. The former instructions did not convey information relevant to the construct or measure of ego development, whereas the latter instructions did convey information relevant for raising ego-level scores. It is argued that the higher scores under the be mature instructions might reflect the respondents’ “optimal” ego level (best functioning), whereas the ego-level score under the standard instructions might reflect their “functional” level (normal functioning).

The Washington University Sentence Completion Test (WUSCT; Loevinger, 1985, 1998b) has been studied for its sensitivity to modified test instructions, such
as the explicit instructions to respond maturely (Jurich & Holt, 1987) or to respond in a thought provoking and complex manner (Blumentritt, Novy, Gaa, & Liberman, 1996). Ego-level scores have been found to be moderately responsive to such instructions, which has been interpreted as an indication of measurement unreliability (e.g., Blumentritt et al., 1996; Jurich & Holt, 1987). The main purpose of this study was to explore whether the responsiveness to modified instructions might be due to measurement unreliability or that it might reflect a genuine developmental phenomenon. Slightly increased ego-level scores under modified instructions may reflect an “optimal” ego level, as opposed to the “functional” level measured under the standard test instructions (cf. Lamborn, Fischer, & Pipp, 1994). The psychometric versus the developmental explanation of mean ego-level differences due to modified instructions was investigated by: (a) a review of the available literature on the effect of modified instructions, and (b) an empirical study of the differential effect of two modified instructions on ego-level scores in children and adolescents (age 9 to 15 years).

A secondary purpose of this study was to investigate the susceptibility to modified instructions of the Sentence Completion Test for Children and Youths (SCT–Y; Westenberg, Treffers, & Drewes, 1998). The SCT–Y is a new version of the WUSCT, developed for measuring ego development in adolescents and children over 8 years (see Loevinger, 1998b). Many questions can be asked about the reliability and validity of a new instrument, and some of these have been addressed. For example, Westenberg, van Strien, and Drewes (2001) found that an oral versus written administration of the SCT–Y did not have a significant impact on ego-level scores, not even in 4th to 6th graders or children with limited reading and writing skills. The question addressed here is whether and to what extent the SCT–Y might be susceptible to certain response sets, such as trying to make a good or mature impression. Indeed, this study is the first one to investigate the effects of modified instructions on ego-level scores in a nonadult sample.

The Effect of Modified Instructions on Ego-Level Scores: A Review

Several studies have been conducted to study the effect of modified instructions on the test–retest stability of the WUSCT. The majority of these studies were aimed at

---

1The Washington University Sentence Completion Test (WUSCT) does include a separate protocol for children and adolescents (Form 2–77; see Loevinger 1998b), but some of the items do not have a formally validated and revised scoring manual, and none of the items or scoring manuals were validated specifically for usage with children and adolescents. For a comparison between the items of Form 2–77 and those of the Sentence Completion Test for Children and Youths (SCT–Y), see Westenberg, Treffers, and Drewes (1998).
raising ego-level scores. These studies will be discussed under the general rubric of “best effort” test instructions. Other studies were aimed at lowering ego-level scores, and will be discussed under the rubric of “worst effort” test instructions.

In some of the best effort studies participants were given information about the ego development construct. In three test–retest studies, Redmore (1976) studied the effect of the instructions to complete the WUSCT as would a person of high ego level, after participants had been given explicit knowledge about the concept of ego development by means of: (a) a 25-min lecture ($N = 70$), (b) a five lecture minicourse ($N = 10$), or (c) a graduate seminar on ego development ($N = 13$). The 25-min lecture yielded a significant average increase of about .5 ego level and the graduate seminar yielded an increase of about .9 ego level; the minicourse produced a small and nonsignificant increase. Blumentritt et al. (1996) studied the effect of gaining knowledge about the highest, integrated ego level, by providing participants a description drawn from Maslow’s concept of self-actualization. On retest the average ego level had increased significantly by approximately .5 ego level ($N = 30$).

Two other best effort instructions encouraged participants to improve their scores in ways conceptually consistent with the ego development construct, but without providing explicit knowledge about the stages of development. Blumentritt et al. (1996) asked participants to complete their retest WUSCT in the “most complex, thought-provoking way that you can. (Give your best response.)” (p. 84). This instruction significantly raised the average score by about .3 ego level ($N = 30$). Jurich and Holt (1987) studied the effect of knowing about the general purpose of the test as being a measure of personal maturity. They did not employ a within-subjects design but used a between-subject design. One group of participants received the standard instructions, the other group was informed about the general purpose of the WUSCT and was asked to respond accordingly: “This is a test of personal maturity: Please complete each sentence in as adult and mature a manner as you can” (p. 187). Both groups were retested to study the comparative stability of ego-level scores under both instructions. At first testing, the average difference between the standard condition and the mature condition was .7 ego level; at retest the difference was .4 ego level. However, due to the between-subject design it is unclear whether these differences in ego-level scores were caused by the different conditions. The number of participants in each group was relatively small ($N = 32$) and the observed differences may have been due to accidental group differences related to ego level (e.g., it was not reported whether both

---

2In fact, the exact magnitude of the ego-level differences were difficult to compute because the reviewed articles were based on the previous ego-level notation (including transitional levels) or did not provide the information needed to compute the exact mean ego-level difference. If needed, the data provided in the literature were recomputed and converted to the current system of ego-level notations (Hy & Loevinger, 1996).
groups differed in age). Yet, an overall difference of .55 ego level is consistent with the effects of the majority of the other best effort conditions.

In sum, providing respondents with explicit or implicit knowledge about the ego development construct, combined with instructions to respond accordingly, proved to have a moderate and significant impact on ego-level scores. Other best effort instructions did not contain explicit information pertaining to the ego development construct and did not have a significant impact on ego-level scores. Redmore (1976) studied the effect of instructing participants to “fake good” at retest, that is to make a highly favorable impression on the tester and the person scoring the test. Two such studies did not yield significant ego-level changes, neither up nor down (testing sessions 1 to 2 weeks apart; Ns = 20, 36). Finally, the best effort study conducted by Weiss, Zilberg, and Genevro (1989) instructed participants to be as candid in completing the sentences at retest as they were the first time, while disregarding their first responses (N = 42). The “be candid” instructions did not have a significant effect on the mean ego-level scores. Apparently, the be candid and fake good instructions did not provide respondents with adequate information on how to maximize their ego-level scores. It may be concluded that the WUSCT is not susceptible to modified instructions conceptually unrelated to the ego development construct.

Three of Redmore’s (1976) test–retest studies were aimed at lowering ego-level scores and provided respondents with “worst effort” instructions. After the first assessment of ego level, one group of students had been exposed to the 25-min lecture on ego development (N = 63), and a second group had participated in the minicourse (N = 10). At retest participants were asked to respond as if they were a person of low ego level. A third group was asked to create a highly unfavorable impression on the person scoring the test, without being exposed to the ego development construct (N = 22). In all three groups, ego-level scores dropped dramatically. Computed across the three studies: 69% of the participants had dropped two or more ego levels, 16% had dropped one ego level, 13% had not changed, and only 2% had climbed one ego level.

The reviewed studies were based on small to moderate numbers of participants, varying from 10 to 70 participants. However, despite this limitation, the findings converged on a general pattern of: (a) relatively large decreases under worst effort instructions, (b) moderate but significant increases under best effort instructions conceptually consistent with the ego development construct, and (c) no gains or losses under best effort instructions conceptually unrelated to ego development. The contrast between the large drop under worst effort instructions versus the moderate raise under some best effort instructions may be explained in terms of asymmetry of comprehension (Loevinger, 1998a): Individuals are expected to understand ego levels below and including their own level, but cannot understand ego levels much beyond their own. Therefore, it was to be expected that participants would readily be able to lower their ego-level scores but would be hard
pressed to raise them. Yet, when instructed in a way conceptually consistent with the ego development construct and measure, participants were able to raise their scores by a modest but statistically significant margin of about .5 ego level (range: .3 to .9). Two explanations for this effect of the best effort instructions will be explored.

EXPLANATION 1: MEASUREMENT UNRELIABILITY

The authors of the reviewed studies argued that the observed ego-level differences on retest might be indicative of the unreliability of the WUSCT, due primarily to the ambiguity of the standard test instructions (Blumentritt et al., 1996; Jurich & Holt, 1987; Redmore, 1976; see also Redmore & Waldman, 1975). The WUSCT instructs respondents to “Complete the following sentences” (Loevinger, 1998b, p. 103). No further guidance is provided as to what the test is measuring. The ambiguity of the test instructions is intentional because it allows respondents to “project” their own frame of reference (Loevinger, 1976). However, Blumentritt et al. and others argued that the undesirable side-effect of open instructions might be the unpredictable influence of unknown response sets, adding to the unreliability of the measure due to an increase of error variance. Moreover, ambiguous instructions would not properly motivate participants to take the test and to take it seriously. Not knowing what is expected of them, respondents would not be induced to give their best effort (Blumentritt et al., 1996). Therefore, the obtained ego-level scores may be lower than what the person would be capable of when properly informed about the general purpose of the test (Jurich & Holt, 1987). To offset the blurring effect of unknown response sets, the “mental state” of the participant would need to be standardized (Cronbach, 1970). Jurich and Holt therefore argued that more straightforward instructions would yield a more reliable and more accurate measure of ego development.

However, from the mere increase of ego-level scores on retest it does not automatically follow that the standard WUSCT would be less reliable than the best effort WUSCT. The WUSCT may be equally reliable under both instructions, at least in terms of the interrater agreement and internal consistency of the items. Indeed, even test–retest stability may be very high, despite of the usage of two different instructions at test and retest. Modified instructions may yield different group means, but may not change the rank ordering of the individuals within the group (i.e., stability in terms of test–retest correlations).

The reviewed studies did not systematically address all aspects of measurement reliability. Except for test–retest stability in terms of mean ego-level scores, piece-meal information is provided on the other aspects of reliability. The scant reliability data do not indicate superior reliability of the best effort WUSCT. Weiss et al. (1989) reported no difference for standard and be candid instructions in terms of
interrater agreement and internal consistency of the items. Ironically, the Jurich and Holt (1987) study, specifically designed to demonstrate the superior reliability of the adult and mature instructions, reported a lower test–retest stability of ego-level scores for the modified instruction as compared to the standard open instructions. The participants in the adult and mature condition declined significantly from an average ego level of 6.0 to an average of 5.8 on retest, \( t = 2.74, p < .01 \), whereas the participants in the standard condition obtained the same average score of 5.3 at test and at retest. This finding suggests that the standard WUSCT would be more reliable than the best effort WUSCT.

However, the absence of a mean difference between test and retest would not by itself be sufficient evidence for measurement reliability either. Although the mean level for a group of participants may not have changed, there may be considerable reordering of the participants in terms of their ego level (i.e., low stability in terms of test–retest correlations). The test–retest stability of test scores need always be looked at in terms of group means and in terms of correlations. None of the reviewed studies has done so, and systematic comparisons in terms of internal consistency or interrater agreement have not been presented either. It thus remains to be seen whether there is any real difference in reliability of the WUSCT under different test instructions.

EXPLANATION 2: OPTIMAL EGO LEVEL

The modest but significant increase of about .5 ego level under some best effort instructions may also reflect an “optimal” ego-level score, as opposed to a “normal” ego-level score under the standard test instructions. The idea of an optimal ego score is akin to Vygotsky’s (1978) notion of a proximal zone of development, in which the participant may be coached to display his or her “possible” level (as opposed to the “actual” level). Some of the best effort instructions may indeed be viewed as a form of coaching toward one’s possible or maximum ego level. This distinction between the actual (uncoached) and possible (coached) level of development is similar to Lamborn et al.’s (1994) distinction between functional and optimal levels. The functional level is the level at which one normally functions, whereas “optimal-level performance usually occurs in situations that provide strong contextual support for high-level functioning, including clear task definitions” (Lamborn, et al., 1994, p. 496). If such contextual support is minimal or absent, participants are expected to perform at their functional level.

The developmental psychologist’s distinction between functional and optimal level resonates with the psychometrician’s distinction between “characteristic” versus “maximum” performance (Jackson, 1993). In personality assessment participants are usually asked to provide their most characteristic or typical response, whereas in ability testing participants are usually asked for their best effort. Trans-
lated to the WUSCT: Under the standard instructions respondents are prompted to reveal their characteristic or functional level (i.e., normal functioning); under some of the best-effort instructions respondents are prompted to reveal their maximum or optimal level (i.e., best functioning).

In other words, the modest but significant increases under some of the best-effort instructions may not be due to measurement error, but may reflect the difference between normal and best functioning. As might have been expected on the basis of ego development theory and other stage models, the optimal level is not far removed from the functional level. People may be coached to increase their ego-level scores to some extent, but cannot be induced to exceed their functional ego level by more than one ego level. The optimal level explanation is corroborated by the lack of effect for the fake good and be candid conditions. These instructions did not contain any direct clues about the purpose of the measure or the content of the ego development construct, and did not yield higher ego-level scores on retest. These instructions did not contain sufficient contextual support for maximum or optimal performance.

The optimal level explanation would be strengthened if the following two requirements were met: First, the instrument should be equally reliable under the standard and modified instructions, in terms of interrater agreement and internal consistency of the items. That would be the first indication that the standard method does not yield less reliable scores. Second, even though participants may obtain slightly higher ego-level scores at retest due to best effort instructions, the test–retest correlation of ego-level scores should still be sufficient (i.e., similar to the test–retest correlation for two standard administrations). Such a finding would demonstrate that all participants would have edged up toward their optimal level while maintaining rank order stability.

STUDY AIMS AND EXPECTATIONS

The general aim was to study the differential effect of two modified instructions on ego-level scores in a sample of 9- to 15-year-old respondents. Ego level was assessed by means of the SCT–Y (Westenberg, Treffers, & Drewes, 1998). One of the modified instructions was taken from Redmore (1976); to make a good or favorable impression on the tester (the fake good condition). The other instructions were taken from Jurich and Holt (1987); to complete the items in as adult and mature a manner as one can (the be mature condition). These two instructions were selected because both are best effort instructions, yet they were expected to have a different effect on ego-level scores. The fake good instructions were expected to have no effect on mean ego-level scores because these instructions contain no valuable clues for maximum performance. To make a good impression on others is not necessarily related to a higher level of development. In contrast, the be mature instructions
were expected to have a small but significant effect on mean ego-level scores. The be mature instructions encourage participants to give their best effort in a way consistent with the ego development construct and measure, without being too explicit about the construct or the measure. Hence, these instructions contain valuable clues for maximum performance (i.e., for displaying one’s optimal level).

The modified test instructions were not expected to have a significant impact on the reliability of the SCT–Y. Whether given by standard, fake good, or be mature instructions, the SCT–Y was expected to display sufficient and similar levels of interrater agreement and internal consistency. Moreover, the test–retest stability in terms of the rank order position of the individuals was also expected to be sufficient. In fact, comparable reliability of the SCT–Y under either test instructions was a precondition for meaningful comparisons of ego-level scores obtained under the different instruction modes.

**METHOD**

**Participants**

The two experiments were conducted in The Netherlands at four separate schools: two elementary schools (Grades 4 to 6) and two high schools (Grades 7 to 9), with 127 participants in the fake good experiment (62 girls, 65 boys) and 128 participants in the be mature experiment (64 girls, 64 boys). The average age in the fake good experiment was 12.5 (SD = 1.77); the average age in the be mature experiment was 12.6 (SD = 1.72). No significant age differences were observed for sex.

**Instrument**

The SCT–Y (Westenberg, Treffers, & Drewes, 1998) was used for measuring ego level. The SCT–Y is a new version of the WUSCT (Hy & Loevinger, 1996; Loevinger, 1985, 1998b). On the basis of a large-scale, multistudy investigation of ego development in adolescents and children over 8 years, it was concluded that Loevinger’s model and measure had to be modified to be appropriate for children and youths (Westenberg, Jonckheer, Treffers, & Drewes, 1998).

The selection of the items for the SCT–Y was guided by seven criteria: maximum overlap with Form 81 of the WUSCT, appropriateness for age 8 to 18 years, comparable forms for boys and girls, optimal variety of item content, face validity (experienced relevance), item validity (item–total correlation), and adequate test length. In comparison with the adult form of the WUSCT (Form 81), 21 items are identical (e.g., When a child will not join in group activities …), 5 items were rephrased (e.g., A girl feels good when … instead of A woman feels good when …),
6 items were added (e.g., My biggest fear ...), and the total number was reduced to 32 items printed on two pages (e.g., A husband has a right to ... was dropped). (For a full description of the item selection procedures and the overlap with the WUSCT, see Westenberg, Treffers, & Drewes, 1998.)

The usage of new or rephrased items necessitated the construction of new scoring manuals. Moreover, usage of the 21 original adult WUSCT items with children and adolescents required modifications as well, particularly with respect to the lowest ego levels: a substantial number of scoring categories had to be constructed or rearranged (see Westenberg, Jonckheer, et al., 1998). These changes in the scoring manual led to changes in the descriptions of the lowest three ego levels. As was anticipated, “a more positive image of the earliest stages emerged” (Loevinger, Carlson, Westenberg, & Lasker, 1998, p. 52). The blatant aggression and the oppositional-defiant attitude associated with the Impulsive ego level of adults were not typical of youths at that level. The Impulsive child or adolescent is characterized by a sense of vulnerability, a dependent coping style, and an appreciation of personally enforced rules. The Self-protective child or adolescent appeared less manipulative and exploitative, was more geared toward self-focused forms of control, displayed a self-sufficient attitude and a live-and-let-live philosophy of life. In contrast with Loevinger, Carlson, et al.’s description, the majority of the Impulsive and Self-protective children and adolescents indicated appreciative relationships with other people. The Conformist adolescent did not display the mindless conformity to externally imposed and specific rules, but appeared to adhere to certain ground rules, such as equality, reciprocity, and communality in relationships. Marginal differences were observed for the Self-aware or higher ego levels. (For a detailed account of the similarities and differences with Loevinger’s description of the earliest ego levels, see Westenberg, Jonckheer, et al., 1998.)

The differences with Loevinger’s (1976, 1997) model and measure for the first three ego levels are attributed to the younger age of our participants and to the greater representation of the lowest ego levels. Loevinger, Hy, & Bobbitt (1998) realized that by using adult samples “one cannot extrapolate very far to ego levels that are underrepresented. There are fewer [Impulsive and Self-protective] subjects than would be optimal to give a fair representation of the possible answers and modes of thought at those levels” (p. 24). Despite the necessary adjustments, however, two cardinal features of the ego development construct were maintained: the same stages in the same order, and individual differences in developmental maturity within age cohorts. Moreover, the presence of child and adolescent versions of the various ego levels support the premise that adult ego levels may be traced to development occurring in childhood and adolescence.

The SCT–Y demonstrated adequate interrater and test–retest reliability and, consistent with the requirement that the SCT–Y should tap a one-dimensional construct, the items of the SCT–Y represent a one factor structure. A first indication of the external validity of the revised instrument and stage descriptions was provided
by the results of research on the relation between ego level and anxiety disorders in a population of children and adolescents referred to an outpatient psychiatric clinic (Westenberg, Siebelink, Warmenhoven, & Treffers, 1999). The two most prevalent and debilitating anxiety disorders in children and adolescents were empirically related to conceptually equivalent ego levels: Separation Anxiety Disorder (SAD) was related to the Impulsive ego level, and the Overanxious Disorder (OAD) was related to the Conformist ego level—with age, sex, IQ, and socioeconomic statistically controlled for. The comorbid condition of SAD and OAD was associated with the Self-protective level.

Experimental Design and Procedure

To be able to document change, a within-subjects test–retest design was used. In other words, participants served as their own control. The SCT–Y was split into two one-page halves: The standard instructions were given for the first half, the modified instructions were given for the second half. The SCT–Y was split into two halves, rather than retesting the entire instrument, to prevent carry-over or practice effects and to prevent a loss of motivation. These factors might dilute the potential effect of the modified instructions. The order of the items was counterbalanced: half of the participants received items 1 to 16 under the standard instructions and items 17 to 32 under the modified instructions, and vice versa for the other half of the participants. This was done to ensure that neither set of items might be more susceptible to the modified instructions.

Standard instructions were the following:

This is a sentence completion test, a list of unfinished sentences. The idea of this test is that you finish the sentences. The list consists of two pages. You will first receive one page of sentences which you’ll be asked to finish, and when you are all ready, you will receive the second page. On this first page, please complete the sentences in any way that you wish.

Written instructions on the test form were: “Complete the following sentences in any way that you wish.” These instructions were taken from WUSCT Form 2–77 (Loevinger, 1998b).

The instructions for the fake good experiment were:

You will now receive the second page of the list. This is also a sentence completion test, a list of unfinished sentences. One thing is different. The first time you were asked to complete the sentences in any way that you wished, but now you are asked to make the best possible impression on anyone reading your sentences. Make sure that other people will think “Wow, that’s a very impressive response,” “that must be a great guy or wonderful gal.”
Written instructions on the test form were: “Complete the following sentences in such a way that you’ll make a very good impression on others.” These instructions were adjusted from the fake good instructions provided by Redmore (1976).

The instructions for the be mature experiment were:

You will now receive the second page of the list. This is also a sentence completion test, a list of unfinished sentences. One thing is different. The first time you were asked to complete the sentences in any way that you wished, but now you are asked to complete the sentences in as adult and mature a manner as you can. Please try to respond even more maturely than you might already have done on the first page, just try to outperform yourself. Please give it your best try.

Written instructions on the test form were: “Complete the following sentences in as adult and mature a manner as you can.” These instructions were taken from Jurich and Holt (1987).

The two administration procedures were carried out in-class in two consecutive 20-min sessions led by a research assistant. The teacher remained present to maintain an adequate level of order and concentration. After the completion of the task, participants were debriefed about the purpose of the experiment.

Data Scoring and Analysis

The SCTY data from both experiments were mixed and scored by the same procedures used to score the WUSCT (Hy, 1998; Hy & Loevinger, 1996): (a) for each respondent the 32 responses were typed into a spreadsheet; (b) the responses were then sorted by item and were put in random order (i.e., all 255 responses to Item 1 were grouped together and randomized); (c) information on the participant and the conditions (standard versus modified, and fake good versus be mature) was hidden from view to allow for blind ratings with respect to age, grade, and experimental condition; (d) each response was rated by two independent and extensively trained raters with the scoring manual for the SCT–Y (see Westenberg, Jonckheer, et al., 1998); (e) the relatively few rating differences between the two raters were resolved by asking a third rater to rate these responses, to resolve the differences by a majority vote; (f) the responses and the ego-level ratings were then resorted to the original protocols, yielding two sets of 16 item ratings for each individual (one set for the standard instructions and another set for the modified instructions); and (g) the frequency distributions of the two sets of item ratings were converted to two total scores. First, the item sum score (ISS) was computed by adding the 16-item ratings for both test halves, yielding a standard and an experimental ISS for each participant. Second, the total protocol rating (TPR) was computed by converting the cumulative frequency distribution of the item ratings to a discrete ego-level score on
the basis of the given rules, yielding a standard and an experimental TPR for each participant. (Before the ISS and TPR were computed the distribution of item ratings was doubled, to arrive at 32-item ratings for the standard and experimental test halves.)

RESULTS

Utility

The first question to be addressed concerned the comparative “utility” of the scoring manual of the SCT–Y for the different instruction modes (standard, fake good, and be mature). Utility is defined in terms of the percentage of responses readily classified by the scoring manual for the SCT–Y. Because the scoring manual was developed on the basis of the standard instructions, responses under the fake good and be mature instructions might not readily be classified by this manual. The results indicated that this was not so. The standard instructions yielded a utility of 95.6% in the fake good experiment and a utility of 96.2% in the be mature experiment. A similar utility level was observed for the experimental conditions: 94.4% of the fake good responses and 95.5% of the be mature responses were classifiable by the scoring manual. The differences between the standard and experimental conditions were not significant according to chi-square analyses. Thus, the utility of the scoring manual is comparable for the different instructions, and an average coverage of about 95% is very high considering the open nature of a sentence completion test. Utility data are not available for the WUSCT.

Reliability

The first aspect of reliability concerned the interrater agreement between the two independent raters. In the fake good experiment, the average interrater agreement across the 32 items was 85.2% for the standard condition (range = 71.7 to 98.3%), and was 83.8% for the fake good condition (range = 68.7 to 97.0%). In the be mature experiment, the average interrater agreement across the 32 items was 87.8% for the standard condition (range = 71.6 to 96.6%), and was 85% for the be mature condition (range = 70.9 to 95.5%). These differences for interrater agreement were not statistically significant. An average interrater reliability of about 85% across the different instructions compares favorably with the level of agreement reported in other studies (e.g., Westenberg & Gjerde, 1999). The disagreements were resolved by a third rater (see the method section).

The second aspect of reliability concerned the internal consistency (i.e., Cronbach’s alpha) of the items under the different instructions. The internal con-
sistency of the SCT–Y could not be computed across all 32 items because each participant completed half of the items under standard instructions and the other half under the experimental instructions. In the fake good experiment, the average internal consistency was .82 under the standard instructions (.83 for Items 1 to 16, .81 for Items 17 to 32 of the SCT–Y), and was .79 under the fake good instructions (for Items 1 to 16 and for Items 17 to 32). In the be mature experiment, the internal consistency was .79 under the standard instructions (for Items 1 to 16 and for Items 17 to 32), and was .82 under the be mature instructions (.83 for Items 1 to 16, .80 for Items 17 to 32). In other words, the internal consistency of the items was comparable across the different conditions, with an average alpha of .80. If corrected for the number of items, an alpha of .80 for 16 items would compare favorably with the alphas reported in the literature for the 36 item WUSCT (see Loevinger, 1998a).

The third and last aspect of reliability concerned the rank order stability of the ego-level scores. In the fake good experiment, ego-level scores obtained on the basis of the standard instructions were highly correlated with the ego-level scores obtained with the fake good instructions: .54 for the TPR and .74 for the ISS ($p < .0001$; the higher correlation for the ISS is probably due to the greater range and variance in ISS scores). In the be mature experiment, ego-level scores obtained on the basis of the standard instructions were highly correlated with the ego-level scores obtained with the be mature instructions: .57 for the TPR and .76 for the ISS ($p < .001$). These correlations indicate that by and large individuals maintain rank order stability across the standard and experimental conditions. Indeed, these correlations are hardly lower than the split-half reliability of the WUSCT administered entirely under the standard instructions. Novy and Francis (1992) reported a split-half correlation of .79 for the ISS, based on two 18-item test halves. (No split-half correlations have been reported for TPR scores from the WUSCT.)

The reliability findings warrant the general conclusion that neither procedure was more or less reliable than the other, at least not in the strict psychometric sense. This conclusion allows for the meaningful comparison of ego-level scores for the standard and experimental conditions.

**Descriptive Statistics**

In the fake good experiment, 72 respondents had obtained the same TPR under the standard and modified instructions (56.7%), 28 had gained one ego level in the fake good condition (22%), 3 had gained two ego levels (2.4%), 21 had dropped one ego level (16.5%), and 3 had dropped 2 ego levels (2.4%). This pattern is consistent with regular measurement error: At retest a similar percentage moves up or down the ego-level scale. In the be mature experiment, 69 respondents had obtained the same TPR under the standard and modified instructions (54%), 41 had obtained a
higher ego-level score in the be mature condition (32%; of which only one respondent had gained 2 ego levels), and 18 had dropped one ego level at retest (14%). This finding is consistent with the optimal level hypothesis: Respondents are most likely to raise their ego-level scores, but are hard pressed to raise their score by more than one ego level.

Mean Level Effects

As expected, the fake good instructions did not yield higher or lower mean ego-level scores in comparison with the standard instructions (see Table 1). The observed differences were not significant according to a repeated measures analysis of variance (ANOVA; i.e., main effect of experimental condition was not significant; analyses were conducted for TPR and ISS separately). No significant interaction effects were observed for the order of the items (Items 1 to 16 first and Items 17 to 32 second, or vice versa), or for sex or school (i.e., elementary vs. high school). No group of participants appeared to benefit from the fake good instruction.

Also as expected, the be mature instructions yielded modest but significantly higher mean ego-level scores in comparison with the standard instructions (see Table 1). According to a repeated measures ANOVA, the within-subjects main effect of instruction mode (standard vs. be mature) on ego-level scores was significant at the .001 level (TPR: $F(1, 127) = 10.31$; ISS: $F(1, 127) = 10.76$). No significant interaction effects were observed for the order of the items, or for sex or school. All increases were well within the maximum increase of one ego level postulated by the optimal level hypothesis.

Consistent with the literature on sex differences in ego development (Cohn, 1991), a main effect was observed for sex: Girls outperformed the boys in both experiments and in all conditions (see Table 1).

Relations to Age

The ego development dimension is expected to be related to the age of the participants and to show individual variability within age cohorts. It was therefore studied whether the different instructions would yield similar relations to age and similar variability within the studied age cohort. In the fake good experiment, the correlations between age and the standard TPR and ISS were .52 and .60, and the correlations between age and the fake good TPR and ISS were also .52 and .60 ($p < .001$). In the be mature experiment, the correlations between age and the standard TPR and ISS were .49 and .60, and the correlations between age and the be mature TPR and ISS were .56 and .65 ($p < .001$). The different correlations for the standard versus be mature conditions were not statistically significant according to Fisher’s
<table>
<thead>
<tr>
<th></th>
<th>Fake Good Experiment</th>
<th>Be Mature Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TPR</td>
<td>ISS</td>
</tr>
<tr>
<td>Total</td>
<td>127 3.28a 3.34 .06 110.41 110.40 .01</td>
<td>128 3.09 3.28 .19 104.83 109.14 4.31</td>
</tr>
<tr>
<td>Item order 1</td>
<td>67 3.24 3.22 -.02 109.07 109.32 -.35</td>
<td>60 3.03 3.18 .15 103.37 108.54 5.17</td>
</tr>
<tr>
<td>Item order 2</td>
<td>60 3.33 3.47 .14 111.90 112.83 .93</td>
<td>68 3.15 3.37 .22 106.12 109.67 3.55</td>
</tr>
<tr>
<td>Boys</td>
<td>65 3.15 3.22 .07 106.80 106.87 -.07</td>
<td>64 2.91 3.02 .11 100.70 104.17 3.47</td>
</tr>
<tr>
<td>Girls</td>
<td>62 3.42 3.47 .05 114.90 113.46 .44</td>
<td>64 3.28 3.55 .27 108.96 114.12 5.16</td>
</tr>
<tr>
<td>Elementary school</td>
<td>60 2.82 2.88 .06 99.62 99.06 .56</td>
<td>62 2.81 3.00 .19 98.34 101.90 3.56</td>
</tr>
<tr>
<td>High school</td>
<td>67 3.70 3.75 .05 120.06 120.12 .06</td>
<td>66 3.36 3.55 .19 100.93 115.94 5.01</td>
</tr>
</tbody>
</table>

**Note.** Underlined values are standard deviations; TPR = total protocol rating; ISS = item sum score; standard = standard instructions; fake good = instructions to make a favorable impression; be mature = instructions to complete the sentences in as adult and mature a manner possible; difference = mean difference between standard and modified ego level scores; item order 1 = Items 1 to 16 at test (standard instructions) and Items 17 to 32 at retest (modified instructions); item order 2 = Items 17 to 32 at test and Items 1 to 16 at retest.

aMean ego-level score; using a scale of ego development where 2 = impulsive, 3 = self-protective, 4 = conformist, 5 = self-aware, and 6 = conscientious.
transformation of $r$ to $z$. In other words, the different conditions did not yield significantly different relations between ego level and age. The observed relations with age are comparable to those observed in studies with the WUSCT (see Cohn, 1998).

Data presented in Table 1 indicate that the variability of the ego-level scores within the studied age cohorts was hardly affected by the specific instructions. For example, in the fake good experiment, the standard deviation of the TPR scores under the standard instructions was .81, and was .77 under the fake good instructions. The variability of the ego-level scores for both instruction modes was also comparable within the narrower age cohorts defined by the elementary versus high school samples (i.e., Grades 4 to 6 vs. Grades 7 to 9). None of the observed differences were significant according to an $F$ ratio test for homogeneity of variance.

**DISCUSSION**

The findings support the optimal level hypothesis: If properly instructed in terms of the concept being measured (be mature instructions), the average respondent significantly raised his or her ego-level scores without exceeding the postulated optimal level (i.e., one ego level up from the standard or functional level). In contrast, if not instructed to alter one’s responses in line with the ego development dimension (fake good instructions), the average respondent did not significantly alter his or her ego-level scores, neither up nor down. These findings with the SCT–Y are consistent with the findings based on the WUSCT reported by Jurich and Holt (1987).
and Redmore (1976). Their findings contributed to the construct validity of the WUSCT, and these findings contribute to the construct validity of the SCT–Y. The WUSCT and the SCT–Y do not appear to be susceptible to a social desirability response set, but are mildly responsive to a response set corresponding to the putative characteristic being measured.

The idea that the ambiguity of the standard instructions would contribute to measurement error and that a more straightforward instruction would improve the reliability of the measure (Blumentritt et al., 1996; Jurich & Holt, 1987), was not supported. The SCT–Y was equally reliable under the standard and the modified instructions. Moreover, regardless of the specific instructions, ego-level scores were equally related to age, sex, and word count, and displayed comparable variability within the studied age cohort of 9- to 15-year-old participants. These findings and those of the reviewed studies suggest that different instructions might yield different but equally meaningful ego-level scores: The standard instructions might yield the respondent’s functional ego level; some best effort instructions might yield the optimal ego level. These findings contribute to our general understanding of the ego development construct. Ego levels are not rigid frames of reference but comprise a developmental range, including a level at which one ordinarily functions and a level that may be reached if properly coached. The latter level may be seen as the “developmental edge” of ego development. Each ego level represents a stable frame of reference, yet at the same time contains the potential and impetus for directed change.

Previous studies had not included a systematic comparison of the effects of different instructions on the reliability of the scoring manual and on the relations to age (e.g., Blumentritt et al., 1996; Redmore, 1976). The only study aimed at demonstrating the superior test–retest reliability of the WUSCT under best effort instructions (Jurich & Holt, 1987) failed to do so. As reported earlier, the findings of their study showed that, on average, respondents in the standard condition had obtained the same average ego level at test and retest (perfect test–retest stability), whereas the average respondent in the be mature condition had obtained significantly lower ego-level scores at retest. Thus, the ego-level scores under the be mature instruction were less reliable, at least with respect to test–retest stability of test scores. Following the optimal level hypothesis, respondents in the Jurich and Holt experiment might have fallen back to their functional level at retest. The be mature instructions might be able to temporarily pull respondents up to their optimal level, but this level is hard to maintain, causing them to fall back to their functional ego level at retest (i.e., the level displayed under standard instructions). In contrast, respondents in the standard condition readily displayed their regular or functional level on both occasions, hence the very high test–retest stability.

In addition to arguing that the be mature instructions would yield more reliable scores, Jurich and Holt (1987) also argued that it would be “ethically desirable … to be direct rather than mysterious or devious about what one is trying to do with a
test, if it is not necessary to be vague or indirect in order to obtain valid measurement” (p. 193). However, asking respondents to complete the sentences in any way they wish is not based on deviousness, but is consistent with the idea that ego level represents our core or functional frame of reference, best elicited by means of open instructions imposing no restraints on the respondents. Moreover, in our long experience with the WUSCT and SCT–Y, none of the respondents ever expressed discomfort about the standard instructions. In contrast, participants are usually intrigued by this unusual and challenging diversion from their daily routine. It remains to be seen whether respondents prefer the restraint of best effort instructions or the freedom allotted by the standard instructions.

Ego-level scores under the different instructions are equally related to the length of the response (word count). The significant relation between ego level and word count has also been reported in other studies (see Loevinger, 1998a). To some extent this relation is inevitable because high ego-level ratings often require longer and more complicated responses. To another extent, however, the relation to word count may also be indicative of the partial confound with verbal fluency and intelligence generally (e.g., Browning & Quinlan, 1985; Newman, Tellegen, & Bouchard, 1998; Westenberg & Block, 1993). The present findings indicated that ego-level scores obtained under the modified instructions were related more strongly to word count than the ego-level scores obtained under the standard instructions. The differences were however not statistically significant, suggesting that neither instruction mode would produce a greater confound of ego-level scores with verbal fluency. Interestingly, the fake good instructions led to a significant decline in word count while not leading to a significant decline of ego-level scores, whereas the be mature instructions did not lead to a significant change in word count but did yield a significant increase of ego-level scores. This suggests that slight fluctuations in word count are not necessarily related to fluctuations in ego-level scores, and vice versa, slightly higher ego-level scores do not always require longer responses.

Although the findings are consistent with the optimal level hypothesis, two questions remain: What is the “real” optimal level? and, which instructions yield the most valid scores? Regarding the first question it is noted that the mean increase was rather small (average increase of .19 ego level). This increase may represent the optimal level under these circumstances (i.e., usage of SCT–Y with children and adolescents, using the be mature instruction), but the optimal level may be higher if different samples or instructions were used. These and previous studies indicate that the optimal level ranges from .2 to .9 beyond the functional level, depending on the particular circumstances and test instructions. In other words, the optimal level depends on the particular context, but is not expected to exceed one ego level under any circumstance.

The second question concerns the comparative validity of the standard versus the best effort instructions. In view of the impressive construct validity of the WUSCT
and the SCT–Y under the standard instructions (see Cohn, 1991; Hauser, 1976; Loevinger, 1979, 1998a; Westenberg, Blasi, & Cohn, 1998), the standard instructions appear the most appropriate way to elicit the respondent’s core or functional ego level. Examination of external correlates, however, should decide which instructions yield the most valid scores vis-à-vis external criteria and circumstances.

ACKNOWLEDGMENTS

The research was supported by a doctoral grant from the Department of Psychology and the Institute for the Study of Education and Human Development, Leiden University, The Netherlands. This research was conducted in partial fulfillment of the requirements for the doctoral degree of Martine J. Drewes. The research assistance of Helma Koen and Anne Marie Troost is gratefully acknowledged. We thank Lawrence D. Cohn and Brian E. Vaughn for their very helpful comments on an earlier version of this article.

REFERENCES


P. Michiel Westenberg
Developmental and Educational Psychology
Faculty of Social and Behavioral Sciences
Wassenaarseweg 52
2300 RB Leiden
The Netherlands.
E-mail: westenberg@fsw.leidenuniv.nl

Received April 17, 2000
Revised August 28, 2000