

New Zealand Psychologist

Volume 8

Number 1

May 1979

Properties of an Inventory to Measure Happiness (and Psychological Health)

Richard Kammann, David Christie, Robyn Irwin, and Graeme Dixon

University of Otago

Adequate attention to the psychology of happiness has been delayed partly by the lack of convincing methods of measurement. This paper describes the development of an inventory to measure happiness consisting of over 100 self-report happy and unhappy sentences and adjectives. Reliability and validity appear to be high, and the test seems promising as a research instrument.

Being "happy" or "unhappy" is taken by most people as an important aspect of life, an index to the goodness of one's situation, an implicit criterion for deciding when to change direction, and an important piece of information about friends and relatives. William James generalized this, saying, "How to gain, how to keep, how to recover happiness is for most men at all times the secret motive of all they do and all they are willing to endure." We interpret James to mean women and men equally.

Paradoxically, psychology has had very little to say about human happiness. Although there has been a steady trickle of isolated articles over the decades, happiness has not emerged as a psychological construct; it has lacked definition, it has received no explanation, it has gained no theoretical standing, and is rarely mentioned in psychology texts. This situation is now beginning to change as major studies have emerged under the rubric of *psychological (or subjective) well-being* (Andrews & Withey, 1976; Bradburn, 1969; Campbell, Converse & Rodgers, 1976).

The tardiness of psychology in attending to happiness is not difficult to explain. The

perhaps necessary scientific detour through a strict behaviourism has left the study of experience in the background for several decades. There also seems to be a latent value judgment that happiness is a slightly luxurious focus of interest, whereas clinical psychology has been ongoing about the serious business of alleviating specific miseries such as anxiety, depression, maladjustment and so on. However, humanistic psychologists in the tradition of Maslow and Carl Rogers argue that a preoccupation with pathology restricts our conceptions of human experience and human potential. Bradburn (1967) specifically justified his research into psychological well-being as an effort to replace the medical model with an approach based on "problems in living" (Szasz, 1961).

A more specific deterrent to the study of happiness has been a set of related assumptions about its non-viability as a construct. Happiness is often assumed to be undefinable, to mean different things to different people, to be a catch-all term for a variety of emotional qualities, to be behaviourally unspecifiable, and to depend upon self-reports which are biased by semantic confusion, immediate mood, social desirability, and defensive perceptions. However recent empirical data shows that these assumptions are not well justified, and that reliable and meaningful self-reports of subjective well-being are entirely possible.

Without the endless and efficient secretarial help of Isabel Campbell and Carol Hunter, the reliable computer assistance provided by Peter Graham and Malcolm McQueen, and the persistent encouragement of Peter Bradshaw and Gay Maxwell, this paper would be even further delayed.

In this paper we report the development of a happiness inventory with potential utility to other researchers. Our first assumption was that we could get a more reliable and valid scale by asking many questions about happiness rather than using a single rating scale. However, we also wanted to discover what people mean when they say they are happy or unhappy, so we developed this inventory around a simple six-step or seven-step happiness rating scale called the criterion item. In this sense we assumed that being happy is what people say it is.

Our assumption that many different questions would measure happiness more accurately than a single criterion item is explained by a simple variance model. We assume that each question we ask has four variance components:

- a. Valid variance measuring the current level of happiness;
- b. Systematic error variance, such as response acquiescence;
- c. Item specific variance associated with the semantics of each item;
- d. Random error variance, such as chance fluctuations in associations with the item, or idiosyncratic interpretations of it.

As we shall see, a typical candidate item for the inventory has an r of about .20-.40 with the criterion item, representing variance components (a) and (b). It also has reliable variance of its own, which can be subjectively illustrated by three example items:

My work gives me a lot of pleasure.

I don't like myself. (-)

I feel close to people around me.

Each of these items taps a different aspect of happiness, thus building up the (a) component of the variance. But each item also has independent associations dealing with work, self, and others. To the extent that these unique components are *uncorrelated* with each other they are random and can be combined with the (d) component, random error variance. By definition, random errors can deviate equally in a positive or negative direction, and tend to sum closer to zero as the number of observations increases.

With a few exceptions, scales for measuring happiness have been single-item *ad hoc* ratings scales, varying from three-step scales up to eleven-step scales. Bradburn (1969)

developed a ten-item scale consisting of five items of positive affect and five items of negative affect. He found the positive affect score was uncorrelated with the negative affect score, and proposed that psychological well-being is the consequence of two independent dimensions. However, this model depends on the assumption that Bradburn's positive and negative affect items are representative for those dimensions, but as the items were arbitrarily chosen this is not necessarily so.

Another multiple item scale is the Index of General Affect developed by Campbell, Converse and Rodgers (1976), consisting of eight semantic differential scales for rating one's life: interesting-boring; enjoyable-miserable; useless-worthwhile; friendly-lonely; full-empty; discouraging-hopeful; disappointing-rewarding; doesn't give me much chance-brings out the best in me.

The reliability of the summary score across these scales, in terms of the alpha coefficient, is .89, and it appears to have reasonable face validity to us. However, the correlation of this Index against two criterion rating scales of happiness and life satisfaction produced r s of .52 and .55 which seems lower than the best that might be achieved.

As part of their methodological *tour de force* on the measurement of well-being, Andrews and Withey (1976) ran a complete intercorrelation matrix on 46 variants of individual happiness ratings scales. A basic finding was that a straightforward seven-step rating scale, ranging from feeling "delighted" to feeling "terrible" or from "very satisfied" with life to "very dissatisfied" emerged as having the highest correlations with all other scales, and as most central to the measure of subjective well-being. Their results give us some assurance that our happiness criterion item is a reasonable one.

The Item Pool and Item Validation

A team of six advanced psychology students was given a starting set of 20 model items generated by the first author which might possibly correlate with happiness or unhappiness. These examples were largely focussed on behaviour dispositions ("I smile and laugh a lot"), postures ("When walking I look at the ground"), and body sensations ("My stomach feels tight"). The team was challenged to

develop as many happiness correlated items as they could over several weeks of meetings and discussions. The only restriction on items is that they had to be *situation-free* as far as possible, not assuming, for example, that the respondent is married or has living parents. However, it was necessary to assume that people have experiences about themselves, about other people, about their relationships with people, about the future and in a few items with "work" or "my work".

The student team was also given access to Bradburn's ten items (which they regularly transformed), to eight "worry" items from Stouffer, and to eleven "anxiety" items from Bradburn (1969).

The team also generated eight heuristic dimensions: happiness-sadness; elation-depression; relaxation-tension; peaceful-anxiety; optimism-pessimism; social acceptance-isolation; extraversion-introversion; vocational satisfaction-dissatisfaction, used as stimuli to generate further items, all of which were debated and re-worded until an informal consensus was achieved. Then the students combed mood adjective check lists for positive and negative adjectives.

The team finally submitted an item pool of 250 sentence items and 187 adjective items, about half of which were intended to reflect positive feelings and half negative feelings, for a total of 437 items. These were then cast into a lengthy questionnaire in which each item could be rated as occurring over "the past few weeks" at one of three levels: *rarely or not at all; some of the time; often or much of the time*. All questionnaires included a six-step happiness criterion item (assigned to a position before or after the whole questionnaire on a random basis) as follows: *very happy; fairly happy; a little happy; a little unhappy; fairly unhappy; very unhappy*, also applied over "the past few weeks".¹

The questionnaire was then administered to 135 university, teachers' college, and polytechnic students, after which the Pearson r was calculated between each item and the criterion, using a 0, 1, 2 scale for item responses, and a 0-5 scale for the criterion item. The arbitrary criterion for considering an item as a possible candidate for the inventory was that the r was significant at the .01 level ($r > .24$) and that it was not significantly skewed at the

.01 level. Of the 437 items tried, 112 items (63 sentences and 49 adjectives) met this criterion.

First Cross Validation

These 112 items were assembled into a new questionnaire for cross validation. Two minor changes were made in format. Each item was now rated on a five-step frequency scale of experience: *not at all; occasionally; some of the time; often; all of the time*. The criterion item was also changed in a largely unsuccessful attempt to get rid of its severe skewness (-1.22) with a new seven-step scale: *extremely happy; very happy; pretty happy; somewhat happy; a little happy; a little unhappy; fairly unhappy*.

This cross validation was part of another study which was concerned with self-peer ratings of happiness, and was administered to 78 people (40 males and 38 females) living in flats in the university area, as tracked down by the experimenter (R.I.). The mean age was 22.1 ($SD = 4.7$) and 40 percent were upper level university students.

Results on these 112 items are shown in Table 1, which summarizes both the original validation study ($N = 135$) and the cross validation ($N = 78$). While we expected to find a regression toward the mean, a shrinkage of item validities, the mean item validity actually increased from an r of .32 in validation to .41 in cross validation.

The frequency scale for each item was assigned values from 0 to 4 allowing us to calculate mean frequency of experience for each item category, yielding scores for each of *positive sentences, negative sentences, positive adjectives, and negative adjectives*. To reduce response acquiescence and method effects belonging with these four types of items, we also derived net scores. *Net sentences* was the mean positive sentence score minus the mean negative sentence score. A parallel operation produced *net adjectives*. Similarly *net all* was defined as the mean response on all positive items (sentences and adjectives) minus the mean of all negative items. (Net scores had a possible range from -4 to +4.) *Net all* is the total inventory score.

¹ Special appreciation is due to the team of six students who patiently generated the 437 items: Jane Berry, Savita Bhana, Helen Dalziel, Anne Marie van Heezik, Gillian Pow, and Julie Sones.

Table 1
Original Validation and Cross Validation of 112 Items:
Pearson *r*s with the Criterion Item

Type of Item	N Items	Original Validation (N = 135)			Cross Validation (N = 78)		
		Mean <i>r</i>	range of <i>r</i>		Mean <i>r</i>	range of <i>r</i>	
Positive Sentences	41	.36	.24	.58	.42	.23	.67
Negative Sentences	22	-.30	-.24	-.46	-.27	.01	-.42
Positive Adjectives	29	.28	.24	.45	.50	.23	.70
Negative Adjectives	20	-.29	-.23	-.43	-.30	-.07	-.51
All Items ^a	112	.32	.23	.58	.41	-.01	.70

^a The result for All Items disregard +/- signs.

The intercorrelations of the subscales are presented in Table 2, along with their correlations with the criterion item. The first point of interest is that the correlations between corresponding sentences and adjectives are high (.91 for positives and .82 for negatives). However the correlation between positive all and negative all (-.72) is not so high. Subsequent studies will show that even this value is an overestimate, but it clearly does not confirm Bradburn's prediction of a zero *r* between positive and negative affect.

The correlation between the overall inventory score (net all) and the happiness criterion item is .73. This compares favourably with the correlations in the vicinity of .55 found by Campbell, Converse, and Rodgers between their eight-item Index of General Affect and

two similar criterion rating scales, in a random national sample of over 2000 American adults. Table 2 also shows that the subscale correlations with the criterion are not much less, indicating that the Otago inventory has more items than it needs.

Examples of some of the most successful sentence items in the original validation and the cross validation are shown in Table 3 along with sentences which did not survive the initial validation on the 135 tertiary students. It will be seen that the successful sentences are typically more general or global in their frame of reference, while sentences attempting to locate specific behavioural dispositions or bodily sensations were usually not successful.

Parallel lists of successful and unsuccessful adjectives are shown in Table 4.

Table 2
Intercorrelations (*r*s) Among Happiness Subscales and the
Criterion Item in the First Cross Validation (N = 78)

Measure	Mean	Mean	Mean	Mean	Mean	Mean	Net	Net	Net	7-step
	Pos. Snt	Pos. Adj.	Pos. All.	Neg. Snt.	Neg. Adj.	Neg. All.	Snt	Adj	All	
Mean Positive Sentences		.91	x	-.73	-.59	-.70	x		x	.63
Mean Positive Adjectives			x	-.74	-.65	-.70		x	x	.67
Mean Positive All		(positives)		-.76	-.61	-.72	x	x	x	.66
Mean Negative Sentences					.82	x	x		x	-.65
Mean Negative Adjectives						x		x	x	-.67
Mean Negative All					(negatives)		x	x	x	-.69
Net Sentences								.91	x	.68
Net Adjectives									x	.74
Net All								(nets)		.73

x Correlation not reported because the score on one variable includes the score (or a subscore) of the other variable (incestuous correlation).

Table 3
Examples of Valid and Invalid Sentences after the First
Cross Validation^a

Valid	Invalid
I feel things are going my way.	I feel like singing.
I wish I could change some parts of my life. (-)	I feel like dancing.
I feel I'm a complete person.	I find it hard to get up in the morning. (-)
I feel that life is meaningful.	When walking I look at the ground. (-)
I feel life isn't worth living. (-)	My stomach feels tight. (-)
I feel confident about the decisions I make.	I am concerned about aches and pains in my body. (-)
I feel loved and trusted.	I cry a lot. (-)
I feel alone. (-)	I feel destructive. (-)
I can't be bothered doing anything. (-)	I feel like giving.
I think clearly and creatively.	I am easily sexually aroused.
I smile and laugh a lot.	I worry about growing old. (-)
I feel free and easy.	I feel adventurous.

^aThe invalid items were eliminated in the first validation study. The (-) sign after an item means that the correlation with the criterion was negative. For valid items, the sign was the same in both validation and cross validation.

Reliability and Validity

A second draw of 129 items was made from the first validation study of which one third were the same as the preceding study, another third were highly valid correlationally but had significant degrees of skewness, and another third were primarily items concerned with

Table 4
Examples of Valid and Invalid Adjectives after the First
Cross Validation^a

Valid	Invalid
happy	scared (-)
unhappy (-)	friendless (-)
glad	amused
miserable (-)	affectionate
insignificant (-)	offended (-)
lively	critical (-)
depressed (-)	weightless
successful	grey (-)
frustrated (-)	sexy
understood	tired (-)
rejected (-)	quick
free	slow (-)

^aSee footnote Table 3.

behaviour dispositions and body sensations which had item validities in the correct direction but below the .01 significance level in original validation. Although only a few of these low-validity items changed their status in this cross validation, we found in check runs that excluding them from the scoring increased the observed correlations only trivially if at all (increments of .00 to .05) presumably because they followed our model of uncorrelated item-specific variances which were self-cancelling. Since the following findings are based on scoring all 129 items, they could be just a fraction conservative.

The 129-item inventory administered twice, two weeks apart, to 45 adult citizens of Tunedin who were members of the Department of Psychology subject panel (mean age 37, $SD = 14$ years). The format of the inventory used the same three-step rating scale for each item, and the same six-step criterion item described in the first validation study. The time-set, as usual, inquired about feelings over "the past few weeks".

Reliability

The split-half reliability was calculated separately for the first administration and the second administration, then averaged using Fisher's Z , and then corrected by the Spearman-Brown formula. We also calculated the two-week retest stability correlations. Both types of reliability are presented for the various scales in Table 5.

As Table 5 indicates, the split-half reliabilities ranged from .92 to .97 among the

Table 5
Reliability of the 129-Item Inventory ($N = 45$)

Scale	Mean Split-Half r^a	2-week Test-Retest r
Positive Sentences	.92	.94
Negative Sentences	.96	.91
Positive Adjectives	.97	.82
Negative Adjectives	.95	.79
Net Sentences	.97	.92
Net Adjectives	.96	.80
Net All	.98	.88
Happiness Item	—	.58

^aMean r for two administrations, then corrected by Spearman-Brown formula for attenuation.

Table 6
*Intercorrelations (rs) Among Happiness Subscales and the
 Criterion Item in the Second Cross Validation (N = 45)*

Measure	Mean Pos. Snt	Mean Pos. Adj.	Mean Pos. All.	Mean Neg. Snt.	Mean Neg. Adj.	Mean Neg. All.	Net Snt	Net Adj	Net All	6-step Happ.
Mean Positive Sentences		.77	x	-.42	-.39	-.42	x		x	.50
Mean Positive Adjectives			x	-.42	-.45	-.44		x	x	.60
Mean Positive All		(positives)		-.45	-.46	-.47	x	x	x	.59
Mean Negative Sentences					.89	x	x		x	-.62
Mean Negative Adjectives						x		x	x	-.68
Mean Negative All					(negatives)		x	x	x	-.68
Net Sentences								.84	x	.66
Net Adjectives									x	.76
Net All								(nets)		.74

x Correlation not reported because the score on one variable includes the score (or a subscore) of the other variable (incestuous correlation).

^a Each correlation is the mean of two correlations obtained from the first and second administrations of the questionnaire.

four basic subscales, and range from .96 to .98 among the three net scores. Although based on a small sample, these results compare favourably with the alpha of .89 obtained by Campbell, Converse, and Rodgers for their eight-item Index of General Affect.

The test-retest reliabilities were also very high, yielding an *r* of .88 for the net all score, indicating that happiness tends to be stable over two weeks. This is our first hint that happiness scores do not reflect mood at the moment. The much lower test-retest stability of .58 for the happiness criterion item illustrates the advantages of the longer inventory.

The two-week stability of .88 for net all also compares favourably with the 15-minute stability observed in a seven-step "life satisfaction" rating scale used by Andrews and Withey (1976, p. 192), yielding test-retest *rs* in the range .61-.71, consistent with the lower reliability in our criterion item.

It is not suggested by the high two-week retest stability that happiness is a personality trait. An eight-month retest stability coefficient of .56 has been observed by Campbell, Converse, and Rodgers with their Index of General Affect; other less reliable measures used by Bradburn and Andrews and Withey produce slightly lower values over 6 to 10 months. This suggests that happiness level is susceptible to slow drifts or infrequent shifts, and lies conceptually between a state and trait. Perhaps it should be likened to physical fitness which may or may not be sustained at any

level, or changed, depending on life style and life situation.

Second Cross Validation

The cross validation with the subject panel again showed an *increase* in average item validity from .22 up to .39. The only hypothesis we can offer for this anti-regression in both cross validations is that there was some noise in the first validation study on 437 items.

The intercorrelations of the subscales shown in Table 6 show a slight deterioration from the first cross validation (Table 2) although this was not usually statistically significant. It was significant, however, for the correlation between positive all and negative all which fell from -.72 in the first cross validation to -.47 in the present sample ($t = 2.07$, $df = 121$, $p < .05$). One of the authors (G.D.) has since confirmed the lower value in a representative sample of 203 Dunedin adults with an *r* of -.45 between positive all and negative all. He also found that the correlation is reliably higher with younger adults (under age 30) than for older adults which is consistent with the drop between Table 2 and Table 6. Thus we come back to a partial independence of positive and negative affect, but still nothing like the orthogonality proposed by Bradburn. We view it as a test method effect requiring the use of net scores.

The correlations between inventory scores and the criterion item were as high (e.g. .74 for net all) in this study as in the first cross

validation, both of which values are limited by the unreliability of the criterion.²

One of the vexing properties of single-item ratings scales is the very reliable appearance of high negative skewness. The criterion item produced a skew of -1.22 in the original validation ($N = 135$), of -.87 in the first cross validation, and of -1.20 in the present sample. These are representative of the skewness values observed by Andrews and Withey (1976, pp. 208-209) for other such scales. However, the skewness of the net all scores in the 129-item inventory is only -.10, although the mean is still consistently above the scale midpoint of zero.

Concurrent and Discriminant Validity

All members of the subject panel routinely complete several standard psychological tests when signing up. Although most of the data on file for the 45 subjects in this study had been collected two years earlier, their records provided one of the authors (D.C.) an opportunity to correlate the 129-item inventory net all scores against some outside tests as shown in Table 7. (Scatterplot inspection revealed no cases of hidden curvilinearity.)

The very low correlations with age and verbal IQ are consistent with previous findings using *ad hoc* rating scales of happiness (see Wilson, 1969, for a review).

The modest correlation of .30 between net all and the EPI lie scale gives some indirect reassurance that the inventory is not seriously contaminated with social desirability. Campbell, Converse, and Rodgers (1976) found an r of only .20 between their measure of subjective well-being and a short form of the Crowne-Marlow Social Desirability Scale, and demonstrated that no correction factor was needed for social desirability.

The high correlation observed with EPI neuroticism (again, net all is superior to the criterion item) is rather surprising considering that Eysenck's purpose was to discover stable factors of personality while our purpose was to develop a uni-factor happiness scale. This is

² The observed r s of .73 and .74 between net all and the criterion item should not be compared with $r = 1.00$ but with the maximum expected r given the unreliability of the criterion. Andrews and Withey's finding of retest r s in the range of .61-.71 for a similar seven-step rating scale provides a median reliability estimate of .66, the square root of which is .81 which is the maximum expected r .

Table 7
Pearson r s between Happiness Inventory or Happiness Criterion Item with Other Variables ($N = 45$)

Variable	Net All	H Criterion
Age	-.17	.08
Raven Matrices	.04	.23
Mill Hill Vocabulary	-.15	.17
EPI Lie Score	.30*	.28
EPI Extraversion	.09	.04
EPI Neuroticism	-.69**	-.41*

** $p < .01$

* $p < .05$

consistent with the hypothesis of a strong negative association between happiness and "maladjustment" or "mental illness". We have subsequently found, in three other samples, that an r of -.70 is representative of the correlation between net all scores and the Beck Depression Inventory.

Memory as a Factor in Test Stability

Although the respondents in the preceding study of two-week retest stability were instructed not to try to remember their responses to the first questionnaire, one of the authors (D.C.) conducted a small study to see if test-retest stability over two weeks could be produced from memory alone. In order to unconfound memory of previous responses with any true continuity of feelings, the subjects in this experiment were asked to copy down the responses of other unknown respondents and without warning were asked two weeks later to recall these responses.

Method

Subjects. The subjects were 137 students at the University of Otago taking first-year Psychology, recruited from six of the laboratory classes. The mean age was 19.5 years with 84 percent falling in age groups 18-21; 69 were male and 68 were female.

Design. The subjects were assigned randomly to six groups in a 2 x 3 design representing two lengths of the Inventory and three time delays for the memory test: immediate; 1-week; 2-weeks. Anticipating the problem of subject-loss over successive laboratory meetings, larger numbers were allocated to the 1- and 2-week conditions than to the immediate condition.

Materials. The long form was the 129-item inventory described in the preceding cross validation study consisting of 67 sentences and 62 adjectives. The short form used 34 sentences and 33 adjectives selected on a subjective matching basis with the items omitted. As before, the instructions on the questionnaire referred to how the subject had been feeling "over the past few weeks" which was expressed for each item in terms of the five-step frequency scale used in the first cross validation ($N = 78$).

Procedure. The subjects were given a blank copy of either the long or short form of the inventory and a corresponding set of numbers varying from 1 to 5 which were the responses of 25 real subjects from a previous study. The subjects were asked to copy the response numbers given by that person into the questionnaire blanks and to read each item as they did so to form an impression of the personality of the other person, and to write a description of him or her (sex not identified) at the end of the copying.

After the appropriate time interval (immediate, 1-week, or 2-weeks) for the condition, subjects were asked to try to recall as accurately as possible the responses they had copied down, again filling out the short or long form of the questionnaire.

Results

The results are expressed in two forms. First is the r between net all scores for responses copied and responses recalled. The second measure is the proportion of exactly correct responses in the recall.

The results are shown in Table 7 where it can be seen that the copy-recall r s of .45-.65 (median = .53) are consistently below the 2-week stability coefficient of .88 for the net all score observed in the previous study.

Table 8
Net All r s and Proportions of Exactly Correct Responses in the Memory Test

Measure	Form	Recall Interval		
		Immed- iate	1-week	2-week
Copy-recall r for Net All	Short	.51	.46	.65
	Long	.62	.56	.45
Proportion Exactly Correct Responses	Short	.47	.36	.35
	Long	.46	.41	.35
N	Short	26	23	21
	Long	24	22	21

Unexpectedly, recall for the long form of the inventory is as high or higher as the short form except after two weeks, and there is no consistent forgetting curve.

The expected forgetting curve does show up in terms of the proportion of exactly correct responses, but again there is no consistent difference between the short and long forms. This might be explained by assuming that subjects "remembered" the responses, not item by item, but by reconstructing the other person from their summary impression of him or her which would be equally effective in the short and long forms. This strategy was in fact mentioned by several subjects.

The Effect of Immediate Mood

If subjects are responding to the happiness inventory in terms of their mood at the moment, then we should find that mood affect scores for "right now" are highly correlated with inventory scores over "the past few weeks". We should further find that varying the time-set in the instructions for the happiness inventory (e.g. the past day, the past week, the past month) should make no difference. Finally, we should find that the test-retest stability of the inventory is the same as the test-retest stability of "right now" mood scores.

The present study (carried out by D.C.) was designed to test these hypotheses.

Method

Seventy-five volunteer University students were randomly assigned to three different time-set instruction conditions for filling out the 129-item inventory: (a) over the past day; (b) over the past week; (c) over the past month. Two weeks later the subjects were asked to fill out the inventory again, with the same time-set instructions.

On both occasions all subjects filled out the Nowlis Mood Adjective Checklist (Nowlis, 1965) for how they were feeling "at the moment you read each word". The items in this scale were rated on a five-step intensity scale: *not at all*; *slightly*; *moderately*; *strongly*; *very strongly*. The 33 items on the Nowlis scale represent 11 factorially distinct affects with three items representing each factor; the factor structure had been previously replicated on a sample of University of Otago students by Mr Tony Egan.

Results

The two-week stability r of the net all scores increased with the time-set condition as follows: day group, .25; week group, .57; month group, .83. As expected, the two week stability r s of the 11 Nowlis mood factors were not different among the three groups, yielding median r s of .33, .39, and .32 respectively, or .35 ($p < .01$) when the data were pooled for the 75 students.

To check on the possibility that these low stabilities were due to irrelevant mood factors, the analysis was repeated using only the five Nowlis factors best correlated with the happiness inventory which were: *affection*, *sadness* (-), *vigour*, *surgency*, and *elation*. Again the median stability r for these five factors did not differ among the three groups (.32, .35, .39) and yielded a median r of .36 ($p < .01$) when the data were pooled.

It is apparent, first, that subjects are responsive to the time-set of the instructions for filling out the happiness inventory, and that the longer the time-set the greater the stability.

It is also apparent that mood "right now" has some stability of its own over two weeks. This can be explained by assuming that mood fluctuates around a baseline of happiness which is characteristic of longer time periods so that a person who is in a good mood one day has an above-chance likelihood of being in a good mood two weeks later.

This last point leads us to expect a significant correlation between mood "right now" and happiness, but this correlation should decline as the time-set of the happiness instruction is increased from a day to a month. We calculated the correlation between each of the five relevant Nowlis factors and net all happiness scores in each of the three groups for each of the two administrations. Averaging the results from the two testings, the median mood-by-happiness r s were .34 in the day group, .33 in the week group, and .26 in the month group. With only 25 subjects in each group, the downward trend with time-set is not significant. (The r s are also not significantly different from zero.)

Summary

The conclusions of these studies can be summarized as a number of separate points.

1. The purpose of the project was to develop a reliable and valid multiple-item inventory of happiness.
2. A highly varied item pool of 437 sentences and adjectives was tried out, from which over 100 items were found to be significantly correlated with a simple happiness rating scale as confirmed in two cross validations.
3. The sentence items with the highest item validities were those which had a global frame of reference such as feeling that life is going well, positive self regard, optimism about the future, a sense of meaningfulness, being loved, feeling free and easy. Sentence items dealing with behavioural patterns or bodily sensations were usually not valid.
4. Bradburn's model of independent dimensions of positive and negative affect was disconfirmed, but the correlation between positive and negative affect, on the order of $-.50$, is below the correlations for other combinations of items.
5. The reliability of the happiness inventory

³ An improved 96-item version of the inventory, including short forms, has been developed under the name *Affectometer 1*. Further information about this test, and procedures for ordering it, are given in Kammann (1979).

is very high in terms of internal consistency and exceeds the best available alternative scales.

6. The concurrent validity of the inventory shows that it is highly associated with measures of neuroticism and depression, but is not importantly associated with the EPI lie scale, age, or verbal intelligence.
7. In terms of stability and concurrent validity, the inventory is superior to a single six-step or seven-step happiness rating scale, such as the criterion item.
8. According to this happiness inventory, using a time-set for feelings over "the past few weeks" happiness has high stability ($r = .88$) over a two week period. This estimate of stability is higher than any previously reported on the basis of shorter scales.
9. It is unlikely that this high two-week stability results from an effort to produce consistent responses on the retest. A test of simple memory for such responses produces a pseudo-stability of only .53 which is apparently derived from an overall impression of happiness level.
10. It has been demonstrated that happiness scores on this inventory are not based upon mood at the time of filling out the questionnaire; on the contrary, it seems likely that current mood is partly a reflection of the characteristic level of happiness prevailing during the current time period.

Although these findings are based on small samples, subsequent studies indicate they are representative of the properties of the inventory. The inventory appears promising for further research toward a psychological understanding of happiness and positive mental health.³

References

- Andrews, F.M., & Withey, S.B. *Social indicators of well-being*. New York: Plenum Press, 1976.
- Bradburn, N.L. *The structure of psychological well-being*. Chicago: Aldine, 1969.
- Campbell, A., Converse, P.E., & Rodgers, W.L. *The quality of American Life: Perceptions, evaluations and satisfactions*. New York: Russell Sage Foundation, 1976.
- Kammann, R. *Sourcebook for Affectometer 1*. Dunedin (P.O. Box 8010): Why Not? Press, 1979.
- Nowlis, V. Research with the mood adjective check list. In S.S. Tomkins, & C.E. Izard (Eds.), *Affect, cognition, and personality*. New York: Springer, 1965.
- Szasz, T.S. *The myth of mental illness: Foundations of a theory of personal conduct*. New York: Harper & Row, 1961.
- Wilson, W.R. Correlates of avowed happiness. *Psychological Bulletin*, 1969, 67, 294-306.