CHAPTER 4: PREPARATORY WORD RATING STUDY

The Word Rating Study had three goals: (a) to select a final item pool for the experiment, (b) to obtain rating information for the to-be-remembered words that could be used as covariates in the experiment, and (c) to explore age-related differences in the perception of the words. The rational for the Word Rating Study was to systematically study the to-be-remembered material *before* conducting the memory experiment. In previous studies, it has not been ensured beforehand that young and older adults have the same perception of the to-be-remembered material. Thus, inconsistent findings in previous studies investigating age-related differences in the positive-negative disparity might be due to age-related differences in the perception of the to-be-remembered material (see section *2.2.4 Potential Reasons for Inconsistent Findings about the Positive-Negative Disparity*). To verify that young and older adults generally agree on the emotional meaning of the to-be-remembered words, the Word Rating Study was conducted.

There were two criteria for selecting the item pool for the main experiment. First, positive, negative, and neutral words should have similar word characteristics on memory-relevant dimensions (i.e., word length, word frequency, imagery). Second, young and older adults should not differ in their ratings of these words. These two requirements are difficult to attain given the large number of word characteristics on which positive, negative, and neutral words could differ and on which young and older adults could differ in their evaluation. These practical concerns gave rise to the second goal of this study. If it is not possible to match positive, negative, and neutral words on all word characteristics, one could at least assess these characteristics to investigate their influence. In this context, an additional advantage of a separate word rating study was that the assessment of word characteristics was completely independent from the memory data and the experimental manipulations. And finally, to my knowledge, there is no study available that has compared young and older adults' perception of emotionally-toned words. On the one hand, an examination of age-related differences in word perception could provide interesting insights into the semantic structure of young and older adults. On the other hand, if major age-related differences exist they would represent a further design challenge for the main experiment to select appropriate to-be-remembered words.

To address these three goals, the Word Rating Study was designed as a preparatory study for the central experiment. For this preparatory study, 200 words were first selected based on eight selection criteria from available rating data from young adults. Then, 24 young and 24 older adults were asked to rate these 200 adjectives on six dimensions: valence, arousal, control, imagery, age-relevance, and self-relevance. These rating dimensions were selected for three reasons: The dimensions point either to the emotional meaning (i.e., valence, arousal, control), to memory relevant characteristics (i.e., imagery, self-relevance), or to age-related stereotypes that may influence the processing of the words differently for young and older adults (i.e., age-relevance).

4.1 METHOD

4.1.1 Participants

4.1.1.1 Sample size and Composition

The sample comprised 24 young (aged 20 to 30 years) and 24 older adults (aged 65 to 76 years) and was stratified by sex. Participants were recruited using two strategies: (a) advertisements in local newspapers in the city of Berlin (Germany), and (b) information from a database of individuals who had participated in previous studies in the Max Planck Institute for Human Development. They were informed that the purpose of the study was to investigate individual differences in the subjective evaluation of words. To this end, the study was called *Wortempfinden* [Feelings about Words]. For the two-hour session, participants received 20 Euro.

4.1.1.2 Sociodemographic Characteristics

Table 4 gives an overview of socio-demographic characteristics of the total sample and of subsamples of young and older adults. Expected differences between the two age groups were found in marital status, $\chi^2_{(3)} = 25.56$, $p < .001$; education, $\chi^2_{(3)} = 34.27$, $p < .001$; and employment status, $\chi^2_{(2)} = 44.00$, $p < .001$. In terms of marital status, as expected older participants were more likely to be married or widowed, whereas younger adults were more likely to be single. Young adults had a least a high school degree and were mostly university students, whereas older adults had a lower secondary education and were retired. Moreover, young adults had received more years of educational training than older adults, $F(1,46) = 8.71$, $p = .005$, $\eta_p^2 = .159$.

4.1.1.3 Intellectual Functioning and Self-Reported Well-Being

To facilitate comparison to other studies, some additional sample characteristics that are typically reported in age comparative research were assessed: self-reported well-being and intellectual functioning. Three items were used to assess life satisfaction, physical health and mental health as indicators for subjective well-being: (a) ”How satisfied are you with your present life?”, (b) “How good is your physical health at present?”, and (c) “How good is your mental health at present?” Responses were given on a five-point scale ranging from *very unsatisfied* (1) to *very satisfied* (5) for life satisfaction and from *very poor* (1) to *excellent* (5) for physical and mental health.

Table 4

Socio-Demographic Characteristics of the Total Sample and for Subsamples of Young and Older Adults

	Total Sample *N* = 48		Young Adults *n* = 24		Older Adults *n* = 24	
Age (in years)						
Mean (SD)			24.3 (2.7)		70.8 (3.3)	
Range			20-30		65-76	
Sex						
Female	24	50.0 %	12	50.0 %	12	50.0 %
Male	24	50.0 %	12	50.0 %	12	50.0 %
Marital Status						
Single	18	37.5 %	17	70.8 %	1	4.2 %
Married, Long-term partnership	21	43.8 %	7	29.2 %	14	58.3 %
Divorced	3	6.3 %			3	12.5 %
Widowed	6	12.5 %			6	25.0 %
Education						
Primary education[a]	3	6.3 %			3	12.5 %
Lower secondary education[b]	11	22.9 %			11	45.8 %
High school[c]	24	50.0 %	22	91.7 %	2	8.3 %
College/University[d]	10	20.8 %	2	8.3 %	8	33.3 %
Years of Education						
Mean (SD)			16.1 (2.2)		13.6 (3.5)	
Range			12-19		7-19	
Employment Status						
Full-time employed	4	8.3 %	2	8.3 %	2	8.3 %
Retired	22	45.8 %			22	91.7 %
Student	22	45.8 %	22	91.7 %		

[a]German: Volks- / Hauptschule. [b]German: Mittlere Reife / Realschulde. [c]German: (Fach-) Abitur. [d]German: Fach- / Hochschulstudium

Consistent with the literature on well-being and aging (e.g., Diener & Suh, 1997; Kunzmann, Little, & Smith, 2000; Larsen, 1978), both age groups reported high levels of life satisfaction and mental health. No significant age differences were found for life satisfaction and mental health. Older adults reported lower physical health but not significantly different from the young adults. Table 5 provides means, standard deviations, and the results of analyses of variance (with age group as between-subjects factor) for these sample characteristics.

Table 5

Sample Characteristics for Subsamples of Young and Older Adults in the Word Rating Study

	Means		Standard Deviations		ANOVA[a]		
	Young	Older	Young	Older	*F*	*p*	η^2
Self-Reported Well-Being							
Life Satisfaction	4.04	4.17	0.75	0.76	0.33	.570	.007
Subjective Physical Health	4.00	3.67	0.66	0.82	2.42	.127	.050
Subjective Mental Health	4.13	4.04	0.68	0.55	0.22	.643	.005
Intellectual Functioning							
Crystallized Intelligence	24.67	21.71	2.90	5.39	5.60	.022	**.109**
Fluid Intelligence	65.63	43.25	12.01	8.36	56.08	<.001	**.549**

Note. Effects in bold are significant at $p < .05$.[a]Degrees of freedom for all *F*-Tests were (1,46).

Two indicators of intellectual functioning were assessed: a vocabulary test for crystallized intelligence and a perceptual speed test for fluid intelligence. Participants completed the Vocabulary and the Digit Symbol Substitution test (DSS) of the HAWIE-III (Tewes, 1991; a German version of the WAIS-R, Wechsler, 1981). Young and older adults differ in both measures of intellectual functioning. Young adults were better in perceptual speed (fluid intelligence) and reported more correct definitions in the verbal knowledge test (crystallized intelligence). The age-related difference in perceptual speed was consistent with the literature on cognitive aging (e.g., Salthouse 1996; Verhaeghen & Salthouse, 1997). Regarding verbal knowledge, past research has often shown that older adults are as good or even better than young adults in tasks measuring crystallized intelligence (e.g., Schaie, 1994). The contrary pattern in the current sample is probably due to both the sample size and the highly educated subsample of young adults. As mentioned above, all young adults had a high school degree, whereas only 10 older adults had a high school degree. The same pattern could be observed in years of education; young adults had on average two years more in formal training than older adults. Taken together, both age groups seem to be positively selected:

Young adults were highly educated and older adults were in good physical health (at least on the level of self-report).[9]

4.1.2 Word Stimuli

In order to select an initial item pool to be rated in the Word Rating Study, a two-step approach was employed. In a first step, a database of rating information was built including as many adjectives as possible. In a second step, the pool of adjectives was systematically reduced to a final item pool of 200 adjectives.

4.1.2.1 Database of Adjectives

To establish a database of adjectives, I collected information about rating data for German adjectives obtained in previous studies. These rating data were predominantly provided by a book of Hager and Hasselhorn (1994), who brought together several German rating studies (see Table A1 in the Appendix for a list of all available rating dimensions). Afterwards, other sources were checked to determine if relevant adjectives were still missing. If this was the case, they were added. These other sources were: (a) emotion adjectives of the PANAS-X (Watson & Clark, 1994; 60 adjectives), (b) emotion adjectives of the MDBF scales (Steyer, Schwenkmezger, Notz, & Eid, 1997; 24 adjectives), (c) marker adjectives for the Five-Factor Model (Goldberg, 1992; 100 adjectives), and (d) adjectives used in a study by Heckhausen, Dixon and Baltes (1989; 358 adjectives). This procedure resulted in a database of 5432 adjectives.

The information about each adjective in the database was provided by different studies and sources, so that some adjectives were rated on many dimensions whereas other adjectives were not rated at all. Moreover, all information were based on ratings by young adults.

4.1.2.2 Selection Process: Eight Control Criteria

Initially, a database of 5432 adjectives was assembled. I reduced this large pool of words on the basis of eight selection criteria: (a) word structure, (b) infrequency, (c) person descriptor, (d) word frequency, (e) word length, (f) clarity, (g) imagery, and (h) relevance as

[9] The primary focus of this dissertation project was on age-related differences in the positive-negative disparity of emotional memory. Sex-related differences were only considered as possible confounding influences for the main interest of this dissertation. Table A2 in the Appendix provides means and standard deviations separately for women and men for all indicators of subjective well-being and intellectual functioning. Table A3 reports analyses of variance including sex of participants as an additional between-subjects factor. These analyses did not reveal any significant differences between men and women.

person descriptor. These criteria ensured a homogeneous item pool for word characteristics known to have some influence on memory performance.

In an initial screening step, I excluded almost all adjectives that (a) consisted of two or three meaningful subwords (e.g., arbeits-wütig [work-happy], dick-bäuchig [potbellied], mutter-seelen-allein [all alone]); (b) were highly infrequent or uncommon (e.g., despektierlich [disrespectful], schlumperig [sloppy], viril [virile]); and (c) could not be used as a description of a person (e.g., endlos [endless], links [left], thematisch [thematic]). This first selection process resulted in a remarkably reduced list of 1412 adjectives.

For the remaining adjectives, word frequencies were obtained from a web-based database of the German language supported by the University of Leipzig (Projekt Deutscher Wortschatz, http://www.wortschatz.uni-leipzig.de/). This corpus contained over 500 million words (August 2004) and is updated continually. This vocabulary database provides information about *word frequencies* and *word frequency classes* (WFC).[10]

In a second screening step, I applied two specific selection criteria: (d) words with low frequency classes (WFC < 7, i.e., high frequent) and high frequency classes (WFC > 17, i.e., very rare) were excluded, and (e) words with less than 4 and more than 12 letters were excluded. These specific selection criteria excluded extreme cases in the distribution of word frequencies and word lengths. This selection step reduced the total number of to-be-considered words to 1046.

Based on the available ratings from previous studies (see Table A1 in the Appendix), a more fuzzy selection process was applied. As mentioned above, ratings were not available for all words. Therefore, appropriate ratings were estimated for some words. In this third step, I focused on three aspects: (f) clarity of meaning, (g) imagery/concreteness, and (h) relevance as personality descriptor (i.e., personality traits or emotion terms). Words with low values in

[10] In addition to word frequencies, word frequency classes (WFC) were considered as an additional measure of occurrence. The distribution of simple word frequencies is highly screwed and follows a function called *Zipf's law* (Zipf, 1935): There are only few very frequent word and numerous very rare words. Word frequency classes are derived by this function by considering the frequency of the word of interest (f_{word}) and the frequency of the most frequent word in the language (f_{der}). In German, the most frequent word is "der" that accounts for approximately 2 to 3 percent of all written text. The formula is: $WFC = \log_2(f_{word} / f_{der})$.

By computing word frequency classes, only the whole-numbered part is taken from the exact result. Frequency classes are in reversed order than frequencies, that means high frequent words have low frequency classes (e.g., WFC = 7) whereas very rare words have high frequency classes (e.g., WFC = 18).

To my knowledge, word frequency classes were never used in studies of memory research. However, using word frequency classes as a measure of occurrence has at least two advantages: First, the distribution of word frequency classes follows approximately a normal distribution. This is important for many inferential test statistics that assume a normal distribution of the considered variables. Non-normal distributions (e.g., the distribution of simple word frequencies) could result in biased estimates of the true parameters. Second, word frequency classes are highly comparable across databases and languages by means of their relative nature. This should foster cross-cultural comparison.

clarity have many different connotations and meanings. These words would result in additional noise in the data and were excluded. Words with low values in imagery or concreteness are difficult to recall. And in compensating for the fact that adjectives, in contrast to nouns, are already difficult to imagine, these words were excluded. The reason to select all adjectives from the personality domain was to ensure that all words share a similar semantic network. This is especially relevant in comparing emotional and neutral adjectives. Many neutral adjectives do not belong to the personality domain (e.g., yellow, long) and these words are probably encoded differently than emotion terms (e.g., aggressive, amused). This step condensed the number of words to 476.

In a final step, I attempted to select approximately the same number of words within each valence category (i.e., positive, neutral, and negative words) for the final item pool of 200 words. Moreover, I tried to select words that should result in similar distributions of word frequencies, word lengths, imagery scores, and arousal scores across valence categories. However, valence ratings as well as imagery and arousal ratings were not available for all words. Moreover, the rating sources differ between words making this process rather difficult.

To check whether participants use the rating dimensions appropriately, I included some words that were actually excluded in earlier steps as treatment checks. These words were related to the rating dimensions and function as marker words (valence: 'angenehm'-'neutral'-'unangenehm', arousal: 'angespannt'-'entspannt', control: 'kontrolliert', imagery: 'subjektiv', and age-relevance: 'alt'-'jung'). If participants understand the meaning of the rating dimension correctly, the rating patterns of these words should match to the rating dimensions. This selection process resulted in a final set of 200 words for the Word Rating Study. Table A4 in the Appendix provides all words with their English translations.

4.1.3 Rating Dimensions

All 200 adjectives were rated on six dimensions: valence, arousal, control, imagery, self-relevance, and age-relevance. The instructions for each dimension were adapted from instructions given by Paivio, Yuille, and Madigan (1968). The exact German instructions for each dimension are provided in Appendix B.

With one exception (i.e., age-relevance), all dimensions were rated on 7-point scales ranging from 1 to 7. For valence, participants were asked to indicate the feeling of pleasantness elicited by each word from *very unpleasant* (1) to *very pleasant* (7). For arousal, participants indicated the feeling of tension elicited by each word from *very relaxed* (1) to *very tensed* (7). For control, participants indicated the feeling of control elicited by each word from *no control* (1) to *high control* (7). For imagery, participants were asked to indicate how easily each word elicited a visual image from *very difficult* (1) to *very easily* (7). For the self-description, each participant indicated how accurate each word describes himself from *not at all accurate* (1) to *very accurate* (7). Age-relevance, in contrast, was rated on a 5-point scale. Participants were asked to indicate whether a word is *very typical for young adults* (1), *more typical for young adults* (2), *neither typical for young nor for older adults* (3), *more typical for older adults* (4), or *very typical for older adults* (5).

4.1.4 Procedure

Participants arrived at the Max-Planck-Institute for Human Development, Berlin in small groups from two to seven persons. Each session consisted of two parts. In the first part, participants completed a booklet about demographic characteristics, a measure of crystallized intelligence (i.e., Vocabulary test), and a measure of fluid intelligence (i.e., DSS). This first part took approximately 30 minutes.

In the second part, participants were introduced to the rating procedure and were asked to complete two booklets: The first booklet contained material for the dimensions of valence, arousal, control, and imagery. The second booklet contained material for the dimensions of age relevance and self-relevance. Across participants, the order of the dimensions was counterbalanced within the booklets (for details, see Table A5 in the Appendix). Each dimension was treated separately in one section of the booklets. Each section contained an instruction page for this dimension followed by eight pages of 25 words for the ratings. The order of words varied within the different rating dimensions. This second part took approximately 90 minutes.

4.1.5 Data Analyses

To check for entry errors, data were entered twice. All variables were checked for missing values, outliers, or impossible values. Moreover, demographic characteristics were checked for logical inconsistencies (e.g., reporting an university degree by also reporting only 8 years of schooling in total). Inconsistent or impossible values were replaced with missing values. For each participant, the suitability of the ratings was checked by means of the included marker adjectives. Generally, ratings were consistent with those expected for these marker adjectives. To analyze the word characteristics, the rating data was reorganized to match the already existing database of adjectives. All analyses were performed on SPSS 11.5.

For each word, an *emotional intensity* score was computed by means of the absolute intensity of positively- and negatively-toned words (for a similar procedure, see Bradley et al., 1992; Buodo et al., 2002). Besides word length and imagery, emotional intensity is one of the best predictors for later word recall (Rubin & Friendly, 1986). For this purpose, the midpoint (4) on the valence dimension was used as reference value. For example, the valence value 3 (somewhat negatively-toned) would result in an emotionality score of 1 (|3-4|=1), whereas the valence score 7 (very positively-toned) would result in an emotionality score of 3 (|7-4|=3).

4.2 RESULTS

In the result section, I focus on three major topics. First, I examine the word characteristics in general, their inter-correlations, and their correlations to ratings of previous studies. These analyses function as a treatment check for the generalizability of the obtained rating data. Second, I examine age-related differences in the perception of these word characteristics. These analyses are informative for the third topic that was the primary goal of the Word Rating Study: the selection of a final item pool of negative, positive, and neutral words for the experiment. Appendix C provides detailed information about the ratings of each word separately for subgroups of young and older adults as well as men and women.

4.2.1 Word Characteristics

4.2.1.1 Marker Adjectives for Six Rating Dimensions

To check, whether participants understood the task and the rating categories employed, I examined the words at the bipolar ends of each dimension. For all six dimensions, Table 6 lists the six words with the highest scores and the six words with the lowest scores. These marker adjectives at the bipolar ends of each dimension indicate that participants treated the rating categories correctly. All words are presented in alphabetical order in Table C1, together with means and standard deviations for each dimension (i.e., valence, arousal, control, imagery, age-relevance, self-relevance).

Words generally associated with a negative-tone (e.g., brutal, verlogen) were rated as unpleasant whereas positively-toned words (e.g., glücklich, gesund) were rated as pleasant. Words that connote a high degree of tension (e.g., aggressive, kämpferisch) were rated higher on this dimension whereas the opposite is true for words involving a feeling of relaxation (e.g., entspannt, zufrieden). A feeling of control was associated to active words (e.g, entschlossen, aktiv) but not to words associated with some degree of helplessness (e.g., hilflos, verwirrt). Easy to imagine words had a very concrete and visible meaning (e.g., alt, häßlich) whereas hardly to imagine words were very abstract (e.g, subjective, neutral). For the self-relevance ratings, nearly all participants indicated that they were very tolerant [tolerant] but not dumb [dumm]. For age-relevance, both marker adjectives young [jung] and old [alt] were rated as one would expect, namely very typical for young adults or very typical for older adults respectively.

Table 6

Adjectives at the Bipolar Ends of each Dimension

Valence		Arousal		Control	
very pleasant (7)		*very tense (7)*		*high control (7)*	
glücklich	6.79	brutal	6.67	entschlossen	6.38
ehrlich	6.56	aggressiv	6.66	konzentriert	6.21
erfreut	6.53	grausam	6.56	aktiv	6.17
gesund	6.52	kämpferisch	6.44	erfolgreich	6.08
einfühlsam	6.52	feindselig	6.35	erfahren	6.02
intelligent	6.48	angeekelt	6.32	interessiert	5.94
• • •		• • •		• • •	
depressiv	1.46	gelassen	1.81	dumm	2.04
aggressiv	1.42	ruhig	1.79	deprimiert	1.98
fies	1.40	angenehm	1.60	zerstreut	1.96
verlogen	1.33	zufrieden	1.54	verwirrt	1.92
grausam	1.10	entspannt	1.44	depressiv	1.85
brutal	1.08	gemütlich	1.44	hilflos	1.55
very unpleasant (1)		*very relaxed (1)*		*low control (1)*	
Imagery		Self-Relevance		Age-Relevance	
easy to imagine (7)		*very typical for oneself (7)*		*very typical for older adults (5)*	
alt	6.48	tolerant	6.23	alt	4.79
hässlich	6.33	ehrlich	6.19	erfahren	4.40
attraktiv	6.32	interessiert	6.10	weise	4.38
traurig	6.30	einfühlsam	6.02	krank	4.33
jung	6.21	friedlich	5.96	einsam	4.25
fröhlich	6.21	treu	5.92	vorsichtig	4.13
• • •		• • •		• • •	
angepaßt	2.94	schuldig	1.73	aktiv	1.77
diskret	2.90	boshaft	1.71	kraftvoll	1.75
neutral	2.72	angeekelt	1.69	ungestüm	1.75
liberal	2.54	grausam	1.38	lebhaft	1.72
normal	2.50	brutal	1.33	spontan	1.65
subjektiv	1.60	dumm	1.31	jung	1.25
hardly to imagine (1)		*not typical for oneself (1)*		*very typical for young adults (1)*	

Note. All scales ranged from 1 to 7. One exception was age-relevance ranging from 1 to 5.

In this context, I would like to mention that words that were rated as very typical for older adults were also rated as both positive and negative in valence. For example, experienced [erfahren] and wise [weise], both rated as very pleasant (M_{erfahren} = 6.15, M_{weise} = 6.38), were rated as the second and third most typical characteristic of older adults. At the fourth and fifth rank, however, ill [krank] and lonely [einsam] emerged that were rated as unpleasant (M_{krank} = 2.08, M_{einsam} = 2.44). This is perhaps also true for ‘young adults’ descriptors given the zero correlation between valence and age-relevance (r = -.06, see section *4.2.1.2 Correlations between Rating Dimensions*).

In sum, the patterns of words that were rated as very high or very low on one dimension were consistent with normative expectations about each dimension. Participants seemed to respond to each dimension in expected ranges.

4.2.1.2 Correlations between Rating Dimensions

Table 7 provides the inter-correlations between rating dimensions, together with the derived scores of emotional intensity and the measures of word frequency, word frequency class, and word length. Additionally, Figure C1 in the Appendix shows a scatter matrix between all six rating dimensions.

Table 7

Correlations between Word Characteristics in the Word Rating Study for all Words (below diagonal, N = 200) and for the Final Item Pool (above diagonal, N = 90)

	1	2	3	4	5	6	7	8	9	10
1. Frequency		**-.85**	.03	.20	.02	-.09	.03	.16	.14	.13
2. Frequency Class	**-.74**		.13	-.16	-.06	-.06	.04	.08	-.19	-.09
3. Length in Letters	**-.15**	**.20**		.05	-.10	.08	.11	-.14	.17	-.14
4. Valence	**.29**	**-.34**	.05		.11	**-.68**	**.65**	.06	**.92**	-.03
5. Emotional Intensity	.05	-.08	-.13	.07		.02	.15	**.25**	.10	-.07
6. Arousal	**-.16**	**.21**	**.15**	**-.62**	.02		-.16	.01	**-.54**	**-.25**
7. Control	**.25**	**-.17**	.13	**.64**	.09	-.09		-.10	**.71**	-.13
8. Imagery	.09	**-.14**	**-.24**	-.06	**.29**	.09	-.14		-.02	**-.30**
9. Self-Relevance	**.24**	**-.26**	**.15**	**.91**	.00	**-.47**	**.70**	**-.15**		-.05
10. Age-Relevance	-.01	-.08	-.05	-.06	**-.15**	**-.29**	-.14	**-.24**	-.04	

Note. Correlations in bold are significant at $p < .05$.

The correlation matrix showed a high correlation between word frequency and word frequency class (r = -.74) documenting that these measures shared a large amount of variance

(~50%). Both measures of word frequency showed similar correlation patterns to other measures. However, the correlations seemed to be a little higher for word frequency class than for the simple word frequency. This was may be due to the superior distribution properties of word frequency classes above word frequencies. Moreover, word frequency (and word frequency class) was slightly related to word length indicating that high frequent words were typically shorter than low frequent words.

Valence was clearly related to word frequency. Positive words were more frequent than negative words. There was no significant correlation between valence and word length. Valence was also not related to the derived score of emotional intensity (*r* = .07). Valence was, however, highly related to the ratings of arousal (*r* = -.62) and control (*r* = .64). Negative words involved a more intense feeling of tension/arousal than positive words. And positive words involved a greater degree of control than negative words. However, the ratings of arousal and control were unrelated (*r* = -.09). In addition, the ratings of valence showed a very high correlation to the ratings of self-relevance (*r* = .91) signifying that the more positive a word was evaluated the more typical it was for the participants. Although I had expected a high correlation between valence and self-relevance ratings, this extremely high correlation was somewhat surprising.

The correlations between valence and imagery and between valence and age-relevance were not significantly different from zero. The zero correlation between valence and age-relevance was, however, an interesting finding. This null effect indicates that some adjectives were more typical for young or more typical for older adults but that these words did not differ in the associated valence. To say it differently, it was not the case that personality characteristics assigned to older adults were more negative than personality characteristics assigned to young adults.

The derived scores of emotional intensity showed few significant correlations to other variables. Specifically, intensity was uncorrelated to arousal ratings. This was unexpected. One would expect a moderate correlation between both measures due to the fact that both measures should contain some information about the intensity of the emotional feeling. Intensity was only related to imagery, that is, very emotional words were easier to imagine than non-emotional words; and intensity was related to age-relevance, that is, very emotional words tended to be more typical for young adults whereas non-emotional words tended to be more typical for older adults.

Another interesting aspect was the high correlation between control and self-relevance ($r = .70$). Personality characteristics that were rated as very typical for oneself were also rated as involving a strong feeling of control.

4.2.1.3 Correlations to Ratings of Previous Studies

To verify the generalizability of the obtained ratings in the Word Rating Study, I compared these ratings with available ratings from previous studies. As mentioned in the method section for the Word Rating Study (see section 4.1.2), I compiled findings from previous rating studies into a word database of ratings (see Table A1 in the Appendix for a complete list of previous rating studies in the word database). In this database, ratings for valence, arousal, control (potency), and imagery were accessible. Moreover, ratings of concreteness that is thought to be highly related to imagery were also available. Ratings of previous studies were based on very different sets of words resulting in different numbers of words overlapping with words used in the present Word Rating Study. For the analyses, I considered only such previous studies that had at least 40 words (20% of all words) in common with the present study. The ratings of all previous studies were based on young adults.

To compare the ratings in the Word Rating Study with ratings from past studies, I contrasted the inter-correlation pattern of valence, arousal, control, imagery, and concreteness ratings with the inter-correlation pattern of corresponding ratings by previous studies. Table 8 provides these inter-correlation patterns for each rating category. The first rating depicted within each category is the rating obtained in the Word Rating Study. The following names indicate the authors of previous studies who had assessed the specific word characteristic. Values depicted in bold represent the correlations between one rating dimension of the Word Rating Study and the corresponding ratings of previous studies. High values document high consistency between the present ratings and past ratings.

For valence, the correlations between the current rating and past ratings were extremely high ($r > .94$) signifying that the primary variable in this dissertation project, the emotional tone of the word material, was measured highly reliable between different studies. This is even more remarkable given that these correlations were based on five different studies with five different subsets of words.

The correlations between arousal measured in the Word Rating Study and arousal measured in past studies were moderately high. The correlations ranged between $r = .30$ and $r = .75$. Obviously, all studies shared some amount of variance, but there was also a substantial amount of variance involved that was specific to each study. Interestingly, all arousal ratings showed large negative correlations to age-relevance in the present study. This correlation indicated that high-arousing personality characteristics were more typical for young adults and low-arousing personality characteristics were more typical for older adults. The moderately high correlations with findings from previous studies suggested that the current method for rating arousal might have differed from previous studies.

Table 8

Correlations between Ratings from the Word Rating Study and from Previous Studies

	1	2	3	4	5	6	7	8	9	10
Valence										
Word Rating Study	.26***	-.35***	.05	**1.00**	.07	-.62***	.64***	-.06	.91***	-.06
Hager et al., 1985[a]	.27**	-.37***	.12	**.97*****	.03	-.61***	.60***	-.03	.91***	-.07
Möller & Hager, 1991[b]	.17	-.19	-.14	**.94*****	.17	-.57***	.55***	.09	.84***	-.01
Ostendorf, 1994[c]	.27**	-.35***	.15	**.95*****	.15	-.56***	.67***	-.02	.92***	.03
Schwibbe et al., 1981[d]	.19	-.24	.03	**.95*****	-.14	-.68***	.62***	-.25	.89***	-.01
Schwibbe et al., 1994[e]	.40***	-.47***	.09	**.97*****	.13	-.57***	.65***	.04	.91***	-.03
Arousal										
Word Rating Study	-.12	.20**	.16*	-.62***	.02	**1.00**	-.09	.09	-.47***	-.29***
Ostendorf, 1994[c]	.07	.00	.14	.33***	.16	**.30*****	.61***	.22*	.42***	-.64***
Schwibbe et al., 1981[d]	-.03	.12	.10	-.29*	.45**	**.75*****	.15	.41**	-.19	-.75***
Schwibbe et al., 1994[e]	.09	-.05	.04	.10	.15	**.42*****	.29*	.28*	.17	-.69***
Control / Potency / Dominance										
Word Rating Study	.25***	-.24***	.13	.64***	.09	-.09	**1.00**	-.14	.70***	-.18
Ostendorf, 1994[c]	.25**	-.24**	.17	.74***	.15	-.20*	**.86*****	.04	.75***	-.21*
Schwibbe et al., 1981[d]	.22	-.15	-.01	.33*	.33*	.29*	**.73*****	.20	.39**	-.59***
Schwibbe et al., 1994[e]	.32**	-.29*	.17	.63***	.24*	.01	**.86*****	.05	.66***	-.26*
Imagery										
Word Rating Study	.09	-.09	-.23***	-.06	.29***	.09	-.14	**1.00**	-.15*	-.24***
Hager et al., 1985[a]	.09	-.10	-.27**	-.11	.14	-.06	-.19*	**.86*****	-.16	-.09
Möller & Hager, 1991[b]	.03	-.02	-.25*	.03	.31*	.07	-.17	**.89*****	-.09	-.23
Wippich & Bredenkamp, 1977[f]	.26	-.14	-.48***	-.24	.40**	.08	-.27	**.88*****	-.32*	-.13
Concreteness										
Hager et al., 1985[a]	.17	-.12	-.24**	-.10	.12	-.05	-.16	**.71*****	-.16	.00
Möller & Hager, 1991[b]	-.01	.02	-.23	-.16	.09	.22	-.25*	**.69*****	-.16	-.16
Wippich & Bredenkamp, 1977[f]	.29*	-.19	-.50***	-.15	.28*	-.07	-.17	**.79*****	-.27	-.01

Note. In bold depicted correlations documenting consistency between previous studies and the Word Rating Study. [a]*N* = 127. [b]*N* = 64. [c]*N* = 136. [d]*N* = 48. [e]*N* = 73. [f]*N* = 50. 1 = Word Frequency. 2 = Word Frequency Class. 3 = Word Length. 4 = Valence. 5 = Emotional Intensity. 6 = Arousal. 7 = Control. 8 = Imagery. 9 = Self-Description. 10 = Age-Relevance.
* $p < .05$. ** $p < .01$. *** $p < .001$.

The obtained ratings of control showed high correlations to ratings of potency or dominance in past studies. Values ranged between $r = .73$ and $r = .86$. The high correlations to previous studies gave support for the appropriateness of changing the dimension to control as a replacement for potency and dominance ratings. In addition, correlation patterns were very similar between studies. Words associated with an intense feeling of control were more frequent, more pleasant, more typical for oneself, and somewhat more typical for young adults.

The scores of imagery in the Word Rating Study were highly related to imagery ratings (.86 ≤ r ≤ .89) and to concreteness ratings of previous studies (.69 ≤ r ≤ .79). The correlations to concreteness ratings were somewhat smaller but still high. Besides, easily to imagine words, in contrast to hardly to imagine words, were shorter, more intense, and involved a somewhat less intense feeling of control.

Taken together, the obtained ratings in the Word Rating Study were generally consistent with ratings from previous studies especially the ratings of valence showed very high correlations between studies. One exception was the dimension of arousal. This dimension showed some differences between studies.

4.2.2 Age-Related Differences in Word Ratings

4.2.2.1 Scatter Plots comparing Ratings of Young and Older Adults

To select an appropriate item pool for young and older adults, word ratings were examined for age-related differences in the perception of word characteristics. To do this, the ratings of young adults were compared with the ratings of older adults. Figure 1 illustrates these correlations as scatter plots. Each dot represents one word. The diagonal stands for perfect agreement between age groups (*r* = 1.00).

Although the correlation between ratings of valence by young and older adults was extremely high (*r* = .91), the scatter plot suggests a number of age-related differences, especially for neutral words. The correlations between ratings of young and older adults were also very high for arousal (*r* = .90) and control (*r* = .92). For both dimensions, the scatter plots did not reveal major discrepancies between young and older adults.

The correlations for imagery (*r* = .80) and age-relevance (*r* = .79) were somewhat smaller. This was probably due to the reduced scale range for these dimensions: Age-relevance was measured with a five-point scale whereas the response scale for imagery was artificially reduced by excluding difficult to imagine words. The scatter plots also suggested fairly consistent ratings of young and older adults for both dimensions. For imagery, the dot at the lower left represents subjective [subjektiv].

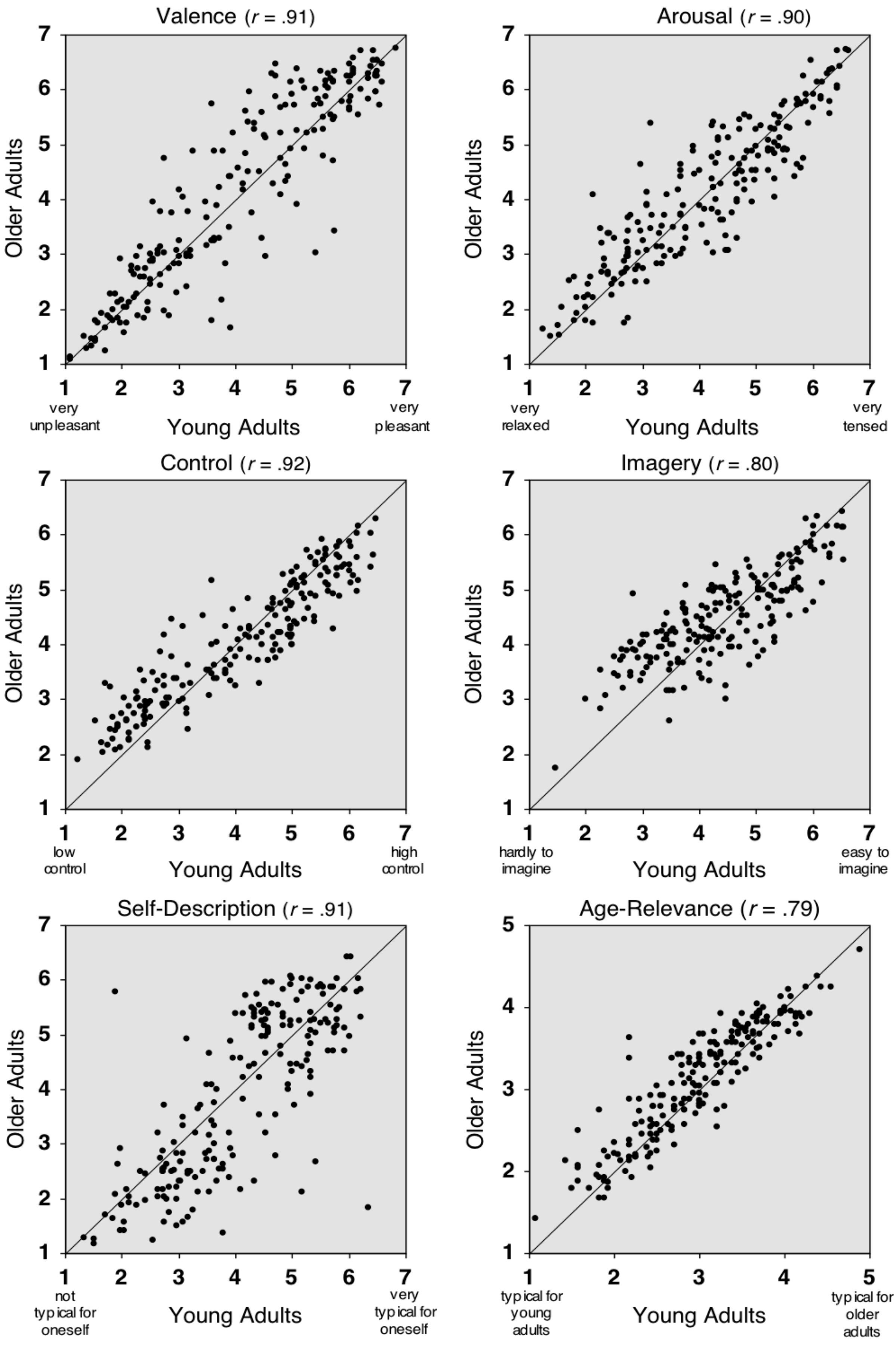

Figure 1 Scatter Plots between Young and Older Adults for each Rating Dimension

Similar to the valence ratings, despite a high correlation between ratings of self-relevance by young and older adults (r = .91), the scatter plot revealed a more disparate pattern from the perfect diagonal. Given the high correlation between valence and self-relevance ratings, the self-relevance pattern resembled the pattern for the valence ratings. The dots at the upper left and at the lower right represent the words old [alt] and young [jung] respectively.

Taken together, the correlation coefficients suggested very high consensus in rating of young and older adults for valence, arousal, control, and self-relevance (r > .90) and also relatively high consensus for imagery and age-relevance (r ≈ .80). Despite the high correlation for valence, the visual inspection of the scatter plot suggested major age-related differences in the valence ratings. This was also true for the dimension of self-relevance but not so much for the other dimensions. The distribution for the self-relevance dimension seemed to resemble the valence dimension.

4.2.2.2 Analyses of Variance for Word Ratings of Young and Older Adults

To address the question of age-related differences in ratings of word characteristics, multivariate and univariate analyses of variance were conducted for each word. For both levels of analyses, the univariate and the multivariate level, age (young vs. old) and sex (men vs. women) functioned as between-subjects factors. Again, sex was included only as a control. This procedure resulted in 6 (dimension) x 200 (words) = 1200 analyses on the univariate level and 200 analyses on the multivariate level. The large number of analyses should result in several significant effects by chance (approximately 5%). However, the significance level for these analyses was not adjusted. The main goal of these analyses was to find words that show age-related differences in the perception of these words and to eliminate those from the final item pool. Thus, not to adjust the significance level was the more conservative procedure in selecting an appropriate item pool for young and older adults.

For each word, a 2 x 2 (Age x Sex) overall MANOVA with the six rating dimensions as dependent variables (i.e., valence, arousal, control, imagery, self-relevance, age-relevance) was carried out. The multivariate analyses of variance revealed for a large number of words significant main effects of age (93 words revealed significant main effects for age, this is 46.5% of all 200 analyses). In contrast, the analyses revealed only a small number of significant main effects for sex (15, 7.5%) and only few significant interactions between age and sex (9, 4.5%). The number of significant main effects of sex and significant interactions were in the range of expected effects by chance.

The univariate analyses were conducted with a 2 x 2 (Age x Sex) ANOVA with age (young vs. old) and sex (men vs. women) as between subjects-factors separately for all six rating dimensions. As expected from the multivariate analyses, the univariate analyses revealed a substantial number of significant main effects of age for valence (62 words with significant main effects of age, this is 31% of 200 words), arousal (42, 21%), control (31, 15.5%), imagery (27, 13.5%), self-relevance (72, 36%), and age-relevance (26, 13%). The total number of significant age-related differences (260, 21.7%) was much higher than the expected number by chance (i.e., 5%). Thus, ratings of word characteristics differ dramatically between young and older adults.

For sex-related differences, the univariate analyses revealed only a small number of significant main effects of sex for valence (27, 13.5%), arousal (8, 4%), control (8, 4%), imagery, (11, 5.5%), self-relevance (23, 11.5%), and age-relevance (15, 7.5%). The total number of significant main effects of sex (92, 7.7%) was hardly larger than the number of significant differences expected by chance (5%). Moreover, the analyses revealed only a small number of significant interactions between age and sex for valence (7, 3.5%), arousal (9, 4.5%), control (12, 6%), imagery (12, 6%), self-relevance (7, 3.5%), and age-relevance (13, 6.5%). The total number of significant interactions (60, 5%) was just the number of significant effects expected by chance. Appendix C provides means and standard deviations for all 200 words and all six rating dimensions. Tables C2 to C7 show ratings of valence, arousal, control, imagery, self-relevance, and age-relevance respectively separately for young and older adults and for men and women. These tables also provide information about effect sizes from univariate analyses of variance.

Taken together, whereas only a small number of words revealed differences in the perception for women and men, many words were rated differently by young and older adults. In particular, age differences were found in the ratings of valence, the primary variable of this dissertation project. To the extent that these differences in perception may also be related to memory performance, this finding has substantial implications for the experimental investigation of age-related differences in memory for emotionally-toned words. If young and older adults differ in their perception of whether a word is more positively- or more negatively-toned, this could have consequences for age-related differences in memory for this word. In sum, the large number of age-related differences found in the perception of the word material points to the necessity of actually assessing whether young and older adults differ in the perception of the to-be-remembered material. This preparatory study has therefore justified.

4.2.3 Selection of an Item Pool for the Experiment

The primary goal of the Word Rating Study was to select an item pool for the experiment. Based on theoretical considerations and pilot work, I made the design decision to use 30 words as to-be-remembered material within one list. Due to the design features of the experiment, it was necessary to select 30 positive, 30 negative, and 30 neutral words in total. The following sections give details about this selection process.

4.2.3.1 Selection Procedure

To select a final item pool of words for the experiment, I used a step-wise approach. In this step-wise selection process, words were excluded based on three criteria: (a) words revealed different valence categories for young and older adults, (b) words were extreme cases in word frequency and word length, and (c) words were phonologically and semantically related.

In the first step, I focused on the primary variable of this project: the valence of the words. To be considered as a positive, negative, or neutral word, I defined ranges on the seven-point scale of valence. To be classified as a negative word, the mean ratings of valence should lay below 2.75; for neutral words between 2.75 and 5.25; and for positive words above 5.25. These ranges were chosen to have approximately the same number of words within each category.[11] One major objective of the Word Rating Study was to ensure that young and older adults show comparable ratings for the final item pool, especially for the emotional tone of these words. To enforce this prerequisite, words had to be rated by both young and older adults as positive, negative, or neutral as defined by the valence ranges. This criterion ensured that both age groups perceived a word as negative, neutral, or positive. If valence ratings of young and older adults indicated different valence categories for the same word, this word was excluded. Following this criterion, 47 words (23.5%) were excluded from further consideration.

In a second step, I narrowed the ranges for word frequencies and word lengths. This was done to make the total set of words more homogeneous. Words were excluded that show either (a) word frequency classes above 16 or below 10, or (b) word lengths below 4 or above 11. This criterion excluded 15 additional words (7.5%).

[11] In an initial attempt, I tried to use stricter ranges for the valence categories (negative < 2.75; 3 < neutral < 5; 5.25 < positive). However, these more restricted ranges resulted in too few words in the neutral valence category.

Finally, the remaining words ($n_{\text{negative}} = 45$, $n_{\text{neutral}} = 47$, $n_{\text{positive}} = 46$) were subjected to a more fuzzy selection process. In this step, I focused on the objective of selecting sets of word that were comparable with regard to word frequencies, word lengths, and imagery scores. It has been shown that at least these three word characteristics are related to memory performance (e.g., Rubin & Friendly, 1986). Therefore, to provide the opportunity to adequately interpret possible memory differences between positive, negative, and neutral words, these words should be matched at least on these characteristics. Prior to the conduct of the Word Rating Study, I planned also to equate positive, negative, and neutral words on perceived arousal level. However, the high correlation between ratings of valence and arousal ($r = -.62$) meant that it was not feasible to pursue this intention.

For the final selection process, I focused on phonological and semantic aspects. Some words were phonological related meaning that they shared identical phonemes (e.g., 'an-ge-spannt'-'ent-spannt') whereas other words were semantically related meaning that they were synonyms or anonyms of each other (e.g., 'stark'-'kraftvoll'). The decision whether words were semantically related or not was aided by a German dictionary of synonyms (Duden, 2004). Words were excluded such that phonological and semantic relations were minimized in the final item pool. Moreover, the selection process took into account the aim of comparable sets of positive, negative, and neutral words with regards to word frequency, word length, and imagery.

As shown by the analyses of age-related differences in the perception of the words, 62 of the 200 words revealed differences for valence ratings of young and older adults. Most of these words were already excluded in step one of the selection process. I tried to exclude the remaining words that show age-related differences in valence. However, it was not doable to exclude actually all words that show age-related differences in valence without breaking the objective of comparable word sets of positive, negative, and neutral words with regards to word frequency, word length, and imagery. I decided to appraise the necessity of comparable word sets of positive, negative, and neutral words as more important than guaranteeing no differences between young and older adults in perceiving the valence of each single word. If sets of positive, negative, and neutral words were not equivalent in view of important variables for memory performance (i.e., word length, imagery), main effects of valence would be difficult to interpret. Main effects of valence as well as interaction effects with valence could be due to these confounding variables. In contrast, age-related differences in the perception of single words could be problematic for the interpretation of significant interactions between age and valence categories. However, one could deal with this problem in follow-up analyses of the main experiment by using subjectively generated valence categories that take into account interindividual differences in the perception of words. Thus, the final item pool showed comparable sets of positive, negative and neutral words with regards to word frequency, word length, and imagery; however, two negative (i.e., frustriert, verärgert), five positive (i.e., ausdauernd, sinnlich, true, umsichtig, zärtlich), and four neutral words (i.e., albern, erschöpft, ironisch, verträumt) in the final item pool revealed age-related

Table 9

Final Selection of Words for the Experiment

Negative Words	Neutral Words	Positive Words
abhängig	abwesend	amüsiert
ängstlich	albern	angeregt
arrogant	bescheiden	aufmerksam
autoritär	ehrgeizig	ausdauernd
brutal	eigenwillig	begeistert
deprimiert	empfindlich	einfühlsam
egoistisch	energisch	entspannt
einfallslos	erschöpft	fröhlich
enttäuscht	erstaunt	gebildet
feige	genügsam	geduldig
feindselig	gesprächig	gemütlich
frustriert	harmlos	geschickt
gehemmt	ironisch	gesellig
geizig	irritiert	glücklich
gelangweilt	listig	gütig
gereizt	moralisch	heiter
hektisch	naiv	höflich
hilflos	resolut	intelligent
krank	scheu	kraftvoll
launisch	schläfrig	kreativ
neidisch	schüchtern	lebhaft
nervös	skeptisch	mitfühlend
stur	still	sanft
träge	überrascht	sinnlich
traurig	ungestüm	tolerant
überheblich	verlegen	treu
unnahbar	verträumt	umsichtig
unruhig	vorsichtig	vergnügt
verärgert	wählerisch	weise
zornig	zaghaft	zärtlich

Note. Table A4 in the Appendix supplies English translations.

differences in perceived valence. Table 9 provides the final item pool of 90 adjectives. Words are listed alphabetically within each valence category (for English translations see Table A4 in the Appendix). Table 7 presents the correlation matrix for this final item pool.

4.2.3.2 Comparison of Negative, Neutral, and Positive Words

To verify that the selected sets of 30 negative, 30 positive, and 30 neutral words did not differ in memory-relevant characteristics, I analyzed the word characteristics for the three valence categories. For each word characteristic, I performed two analyses of variance: one analyses comparing negative, positive, and neutral words and one analyses comparing only negative and positive words. Thus, valence category was a between-words factor with three levels (positive vs. negative vs. neutral) in one analysis and with two levels (positive vs. negative) in the other analysis. This was done to uncover overall differences between words in different valence categories and especially differences between positive and negative words.

As dependent variables, I used the objective measures of word frequency, word frequency class, and word length, the subjective ratings of valence, arousal, control, imagery, self-relevance, and age-relevance, and also the derived scores of emotional intensity. For these word characteristics, Table 10 provides means and standard deviations separately for the 30 negative, 30 neutral, and 30 positive words. Table 11 provides the corresponding analyses of variance.

The objective measures of word frequency, word frequency class, and word length did not significantly differ between positive, negative, and neutral words. Moreover, the corresponding effect sizes were very low ($\eta^2 \leq .03$) demonstrating the realization of matched word sets with regard to these objective word characteristics.

The subjective rating data showed a more disparate profile between valence categories. Positive, negative, and neutral words differed with regards to valence. This should be the case due to the simple fact that I had selected words in the different valence categories according to their valence scores. Similarly, positive, negative, and neutral words differed with regards to arousal, control, and self-relevance. As mentioned in section *4.2.1.2 Correlations between Rating Dimensions*, these rating dimensions were highly correlated to the ratings of valence (i.e., arousal: r = -.62, control: r = .64, self-relevance: r = .91) making it practically impossible to equate positive, negative, and neutral words on these dimensions. For the experimental investigation of the phenomena of emotional memory, this is maybe a problem for the interpretation of potential differences between valence categories.

Table 10

Word Ratings for the Final Item Pool of 30 Negative (N), 30 Neutral (O), and 30 Positive Words (P)

	Means			Standard Deviations		
	N	O	P	N	O	P
Objective Measures						
Frequency	6.18	7.88	9.46	6.87	8.97	10.92
Frequency Class	12.83	12.70	12.33	1.58	1.78	1.81
Length	7.87	8.30	8.03	1.98	1.78	1.71
Subjective Ratings						
Valence	2.09	3.94	6.07	0.36	0.66	0.34
Arousal	5.20	3.87	3.01	0.99	1.06	0.94
Control	3.06	3.74	4.82	0.95	1.17	0.58
Imagery	4.68	4.30	4.78	0.78	0.74	0.88
Self-Relevance	2.56	3.79	5.38	0.47	0.96	0.42
Age-Relevance	3.06	3.13	2.98	0.58	0.64	0.69
Derived Score based on Valence						
Emotional Intensity	1.91	0.58	2.07	0.36	0.30	0.34

Note. N = Negative Words. P = Positive Words. O = Neutral Words.

Table 11

Analyses of Variance for the Final Item Pool of 30 Negative, 30 Neutral, and 30 Positive Words

	negative vs. positive vs. neutral			negative vs. positive		
	F[a]	*p*	η^2	*F*[b]	*p*	η^2
Objective Measures						
Frequency	0.98	.379	.022	1.94	.169	.032
Frequency Class	0.68	.512	.015	1.30	.258	.022
Length	0.43	.653	.010	0.12	.728	.002
Subjective Ratings						
Valence	517.54	<.001	**.922**	1927.25	<.001	**.971**
Arousal	36.27	<.001	**.456**	76.85	<.001	**.570**
Control	27.16	<.001	**.384**	74.98	<.001	**.564**
Imagery	2.96	.057	.064	0.23	.637	.004
Self-Relevance	137.66	<.001	**.760**	601.71	<.001	**.912**
Age-Relevance	0.41	.662	.009	0.25	.619	.004
Derived Score based on Valence						
Emotional Intensity	176.32	<.001	**.802**	3.16	.081	.052

Note. Effect sizes in bold were significant at p < .05. [a]Degrees of freedom for all F-values were (2,87). [b]Degrees of freedom for all F-values were (1,58).

Memory differences between positive, negative, and neutral words may be due to these confounding characteristics of the word material. For this reason, the rating data was used as covariates in follow-up analyses of the memory data to examine the influence of these confounds in the experiment.

The scores of emotional intensity, the derived measure from ratings of valence, showed a significant difference between valence categories. This difference, however, was due to the very low emotional intensity score of the neutral category. Positive and negative words did not differ in their intensity. Words in the different valence categories did as well not differ on ratings of imagery and on ratings of age-relevance. On the one hand, this demonstrates that the "fuzzy" procedure of selecting equally imaginable words for each valence category worked very well. On the other hand, it demonstrates that the selected sets of words did not differ with regard to the elicited aging-stereotype. Thus, the sets of positive, negative, and neutral words very equally imaginable and did not involve specific attributes that were stereotypical more relevant for young or older adults.

Overall, across valence categories, words were matched for imagery, age-relevance, word frequency, word frequency class, and word length. Moreover, positive and negative words were matched on emotional intensity. Words were not matched on arousal, control, and self-relevance.

4.2.3.3 Age-Related Differences in the Final Item Pool

The previous section investigated general differences in word characteristics of the selected item pool of positive, negative, and neutral words. In this section, I focus on potential differences in ratings of young and older adults of the final item pool.

Separately for both age groups, Table 12 gives means and standard deviations for word characteristics of positive, negative, and neutral words. For each dimension, Table 13 provides two corresponding analyses of variance: one 2 x 3 (Age x Valence) mixed analyses of variance with age (young vs. old) as within-words factor and valence category (negative vs. neutral vs. positive) as between-words factor and a similar 2 x 2 (Age x Valence) mixed analyses of variance comparing only positive and negative words. In the previous section, I considered differences between valence categories in general. Significant effects in the previous section were also significant in the present analyses. Thus, I will not discuss these effects again.

The analyses revealed significant main effects of age for self-relevance and age-relevance. Young adults rated all words as being more typical in describing themselves than older adults did. In contrast, older adults rated all words as slightly more typical for older adults than for younger adults.

Table 12

Word Ratings of the Final Item Pool by Age Group and Valence Category

	Means						Standard Deviations					
	Young Adults			Older Adults			Young Adults			Older Adults		
	N	O	P	N	O	P	N	O	P	N	O	P
Valence	2.08	4.00	6.04	2.10	3.88	6.09	0.38	0.70	0.40	0.40	0.75	0.38
Arousal	5.34	3.89	2.84	5.05	3.86	3.20	1.00	1.09	0.97	1.02	1.11	1.00
Control	2.93	3.72	5.02	3.19	3.76	4.64	1.18	1.34	0.62	0.79	1.04	0.60
Imagery	4.74	4.28	4.68	4.61	4.32	4.87	0.90	0.94	1.04	0.76	0.69	0.81
Self-Relevance	2.83	4.00	5.45	2.28	3.57	5.32	0.59	0.92	0.49	0.50	1.19	0.49
Age-Relevance	3.02	3.11	2.86	3.10	3.14	3.09	0.58	0.71	0.70	0.61	0.63	0.72
Emo. Intensity	1.92	0.59	2.04	1.89	0.66	2.09	0.38	0.35	0.40	0.40	0.35	0.38

Note. N = Negative words. O = Neutral words. P = Positive words.

Table 13

Analyses of Variance for the Final Item Pool by Age Group and Valence Category

	Age Group			Valence Category			Age x Valence		
	F	*p*	η_p^2	*F*	*p*	η_p^2	*F*	*p*	η_p^2
Comparing Negative vs. Positive vs. Neutral Words[a]									
Valence	0.14	.709	.002	517.28	<.001	**.922**	1.39	.254	.031
Arousal	0.05	.827	.001	36.39	<.001	**.455**	11.07	<.001	**.203**
Control	0.20	.653	.002	27.17	<.001	**.384**	12.48	<.001	**.223**
Imagery	0.22	.636	.003	2.97	.057	.064	1.96	.147	.043
Self-Relevance	26.18	<.001	**.231**	137.78	<.001	**.760**	3.08	.051	.066
Age-Relevance	11.50	.001	**.117**	0.41	.663	.009	2.57	.082	.056
Emotional Intensity	0.87	.355	.010	174.64	<.001	**.801**	0.51	.605	.011
Comparing Negative vs. Positive Words[b]									
Valence	0.69	.410	.012	1927.12	<.001	**.971**	0.17	.682	.003
Arousal	0.28	.596	.005	76.66	<.001	**.569**	23.85	<.001	**.291**
Control	0.78	.380	.013	75.02	<.001	**.564**	23.64	<.001	**.290**
Imagery	0.14	.714	.002	0.22	.639	.004	4.68	.035	**.075**
Self-Relevance	24.23	<.001	**.295**	601.53	<.001	**.912**	9.57	.003	**.142**
Age-Relevance	15.84	<.001	**.214**	0.25	.619	.004	3.20	.079	.052
Emotional Intensity	0.17	.682	.003	3.15	.081	.052	0.69	.410	.012

Note. Effect sizes in bold were significant at p < .05. [a]For the F-values, degrees of freedom for the effects of age groups were (1,87) and for the effects of valence and Age x Valence were (2,87). [b]Degrees of freedom for all F-values were (1,58).

For valence, the primary variable of interest in this dissertation project, no significant main or interaction effects of age were found. Both age groups perceived the selected set of words as equally positive, negative and neutral. Thus, despite a few age-related differences in perceived valence on the level of single words, no age-related differences in perceived valence were apparent on the group level of these words.

Significant interaction effects between age and valence were found for arousal, control, imagery, and self-relevance. For arousal, young adults rated the negative words as more arousing and the positive words as less arousing than older adults. In the same way for control, young adults rated positive words as involving a more intense feeling of control and negative words as involving a less intense feeling of control than older adults did. For arousal and control, both age groups indicated the same values for the neutral category signifying that young adults probably use a wider range of values in scoring these dimensions.

The significant interaction for imagery in the comparison of positive and negative words indicated that young adults rated negative words as somewhat easier to imagine than older adults did, whereas older adults rated positive words as somewhat easier to imagine than young adults did. For the memory experiment, this could be a serious problem for interpreting potential interactions between valence and age groups. If imagery is the driving force behind memory differences, older adults would be expected to show slightly better memory for positive than for negative words whereas young adults should show slightly better memory for negative than for positive words. However, the effect was rather small suggesting that the impact of these differences was quite limited.

For self-relevance, the significant interaction revealed that over and above the main effect of age young adults rated all words as more typical in describing themselves young adults rated negative words, relative to positive words, as even more typical for themselves as older adults did.

4.3 DISCUSSION

This study was devised to address in part a criticism that can be made against previous studies, namely the selection of the to-be-remembered material: Previous studies had not verified beforehand that young and older adults did not differ in their subjective perception of the to-be-remembered material. Thus, age-related differences in the positive-negative disparity of emotional memory might be due to age-related or cohort-related differences in the subjectively perceived valence of the memory material.

To address this query *before* the experiment, the Word Rating Study was conducted to select an appropriate item pool of 30 positive, 30 negative and 30 neutral words for the central experiment. In this preparatory study, young and older participants were asked to rate 200 common adjectives on six dimensions: valence, arousal, control, imagery, self-relevance, and age-relevance. Based on the ratings of valence, a measure of emotional intensity was computed as the absolute deviation from the neutral midpoint (see Bradley et al., 1992, Rubin & Friendly, 1986). In addition to these subjective measures of word characteristics, more objective measures were acquired for each word: word frequency, word frequency class, and word length. Based on these subjective and objective word characteristics, the final item pool for the experiment was composed.

The main goal of the Word Rating Study was the selection of an appropriate item pool for the central experiment. The discussion is organized around the three major themes to attain this goal: The first section deals with general word characteristics and their interrelations; the second section discusses age-related differences in these word characteristics; and finally, the third section discusses the selection process and its advantages and disadvantages.

4.3.1 The Generalizability of the Rating Data in the Word Rating Study

The first part in the analyses was intended to examine the construct validity of the assessed dimensions. In a first step, adjectives at the bipolar ends of each dimension were inspected. This was done to check whether participants understood the instruction for each dimension correctly. The patterns of words that were rated as very high or very low on one dimension were consistent with normative expectations about each dimension. One dimension that was very instructive in this regard was age-relevance: The trait most typical for young adults was young [jung] and the trait most typical for older adults was old [alt].

In a second step, the inter-correlation matrix was examined. In general, the correlations were consistent with expectations. For example, positive words were more frequent than negative words. This finding is consistent with the literature on word ratings (e.g., Ortony et al., 1987). Nevertheless, there were two findings worth mentioning: First, there was a very high correlation between ratings of valence and self-relevance indicating that all positive words were very typical and all negative words were very untypical for all participants' self-concept. Second, emotional intensity, the derived intensity score from the valence ratings, was not significantly correlated to arousal. Thus, both dimensions were independent from each other.

In a final step, the validity of the assessed constructs was verified by ratings from previous word rating studies with young adults (for a list of available rating studies, see Table A1 in the Appendix A). From these previous studies, ratings of valence, arousal, control, imagery, and concreteness were available for reasonable subsets of the 200 words used in the present Word Rating Study. The correlations between dimensions in the Word Rating Study and corresponding dimensions from previous studies were consistently high. In particular, the primary dimension of this dissertation project, the valence of the considered words, showed very high correlations with previous ratings. Similarly, control and imagery of the present Word Rating Study showed high correlations to corresponding dimensions from previous studies. Thus, the validity of these dimensions was consistently high. One exception was the arousal dimension that showed only moderate to high correlations to arousal ratings from previous word rating studies. One reason for this pattern might be that some previous word rating studies used several marker adjectives to label the bipolar ends of their dimensions. This procedure of using multiple labels for the end points is problematic for the reason that some studies used the same word pair (i.e., “active-passive”) to label the arousal and control dimensions. Thus, by using similar labels, ratings of arousal and control have to be correlated. To avoid these confounded ratings, the present Word Rating Study used only one word pair to label each dimension. This might explain the only moderate correlation between the present and previous arousal ratings.

In sum, the assessment of word characteristics in the Word Rating Study was successful. First, adjectives showed expected means on the six rating dimensions. Second, rating dimensions showed expected correlations between rating dimensions. And finally, the present ratings showed high consistency with previous ratings. One exception is the arousal dimension that showed only moderate correlations with previous ratings.

4.3.2 Age-Related Differences in the Evaluation of Emotionally-Toned Words

As reviewed above (see section *2.2 Age-related Differences in Emotional Memory: Empirical Findings from Experimental Approaches*), findings across previous experiments investigating age-related differences in the positive-negative disparity of emotional memory are inconsistent. One potential reason for these inconsistencies could be age-related differences in the perception of the to-be-remembered material. If the emotional tone (i.e., valence) of the to-be-remembered material is related to memory performance and if age-related differences in the perception of the emotional tone exists, this might lead to different memory pattern for young and older adults. Unfortunately, previous studies have selected the to-be-remembered material based on ratings by young adults neglecting potential age-related or cohort-related shifts in the perception of the to-be-remembered material. The Word Rating Study was designed to respond to this potential confounding factor in the investigation of age-related differences in remembering emotionally-toned material.

Indeed, the Word Rating Study revealed major age-related differences in the subjective evaluation of the word material. In particular, young and older adults showed for a relative large number of words significant mean level differences for all six rating dimensions. In particular, young and older adults revealed significant mean level differences in valence, arousal, control, imagery, self-relevance, and age relevance in 31%, 21%, 15.5%, 13.5%, 36%, and 13% of the 200 words, respectively. In contrast, the Word Rating Study did not reveal a major influence of sex-related differences in the perception of the word material. The number of sex-related differences was in the range of expected values by chance signifying that men and women for the most part agree on the emotional meaning of the word material. Despite the significant number of age-related differences, both age groups revealed high consensus on the rank order of the words as evident by the very high correlations between young and older adults. The correlations between ratings of young and older adults were very high for valence, arousal, control, and self-relevance (i.e., r = .90) and even high for imagery and age-relevance (i.e., r = .80). The somewhat reduced correlations for imagery and age-relevance were most likely due to a condensed range for both scales.

In sum, the Word Rating Study revealed a large number of age-related differences in the emotional evaluation of the words. This finding emphasizes the need for a systematic investigation of age-related differences in the emotional meaning of the to-be-remembered material beforehand. Moreover, this finding gives some support for the idea that age-related differences in the emotional evaluation of the to-be-remembered material are also likely in previous studies. (In the next section, I discuss consequences of differences in word characteristics for positive, negative and neutral words. Some of these consequences might also apply to previous studies. However, previous studies did not report specific characteristics of their memory material.) These potential differences might be one factor in explaining inconsistent findings across experiments. For the present experiment, such words were selected that showed as far as possible high consensus between age groups.

4.3.3 The Selection Process for the Final Item Pool

The primary goal of the word rating study was to select an appropriate item pool of to-be-remembered words for the experiment. This item pool should ideally show two features: First, positive, negative, and neutral words show similar characteristics. Second, young and older adults agree on the emotional evaluation of these words. The selection procedure was partially successful in attaining these features.

Regarding the first point, the selected sets of positive, negative, and neutral words did not differ with regards to word frequency, word frequency class, word length, imagery, and age-relevance. Moreover, sets of positive and negative words were also matched on emotional intensity. With the exception of age-relevance, all other characteristics have been repeatedly shown to be highly relevant for remembering words (e.g., Rubin & Friendly, 1986). Moreover, the dimensions of word frequency, word length, and imagery are typically used to match word sets in memory experiments. From this perspective, there is little reason to assume that any valence category (i.e., sets of positive, negative, and neutral words) has a predetermined memory advantage above other categories.

The sets of positive, negative and neutral words did differ, however, with regards to arousal, control, and self-relevance. Given the total number of 200 words, it was simply not possible to match positive, negative, and neutral words as well on these dimensions. The numbers of positive and negative words that show overlapping values in arousal or in control were just to small to generate matched sets. Moreover, these sets would not be matched on word frequency and word length resulting in even greater discrepancies between word sets. Self-relevance and valence were so highly correlated that they were practically interchangeable. Given that valence was the primary variable to categorize words into negative, positive, and neutral word, these word sets differed as well in self-relevance. Actually, the sets of positive and negative words did not show any overlap in self-relevance.

What are the potential consequences of this partial imbalance between positive, negative, and neutral words? If arousal drives superior recall performance for words, negative words would be remembered better than neutral words that would be remembered better than positive words. Thus, memory differences between positive, negative, and neutral words might be influenced by differences in arousal level. One has to acknowledge, however, that the elicited levels of arousal are very low in comparison to the possible range. For example, the IAPS contains high-arousing pictures of burned faces and erotic scenes that are different in the experiential quality from high-arousing words. Thus, it is not clear whether high-arousing words show a memory advantage over low-arousing words. For control, to my knowledge, no specific relation to memory performance has been suggested. However, the dimension of control was assessed to cover the emotional aspects of the word material and not so much the memory-relevant aspects. This might indicate that the impact of control on remembering words is small or negligible. But I am not aware of any study that has actually tested the impact of the dimension of control on remembering words. In the present context, control was assessed as a more exploratory dimension.

For self-relevance, there is some research showing that the instruction to encode to-be-remembered words in terms of the self enhances memory for these words (e.g., Czienskowski & Giljohann, 2002). However, this self-reference effect is an effect of the instruction and not an effect of the relevance of the words to the self. Thus, it is not the case that the self-reference effect enhances memory only for relevant words but rather for all words. In terms of the present experiment, the questions would be whether participants automatically use the self to encode to-be-remembered material and whether the degree of agreement between the semantic meaning of the word and the subjective representation of the participants' self provides memory-relevant information over and above the emotional connotation. To my knowledge, there is no empirical study available that could inform an answer to both questions. As it stands, however, it was not possible to match positive, negative, and neutral words on arousal, control, and self-relevance. In the present dissertation project, I attempted to address this problem by using these word characteristics as covariates in follow-up analyses (see *5.2.7 Follow-up Analyses on Recallability*). The central experiment was set up in such a way to actually examine the impact of word characteristics in recalling words.

The second desired feature of the final item pool was that young and older adults generally agreed on the emotional connotation of the to-be-remembered words. Young and older adults reported similar values for valence and emotional intensity for sets of positive, negative, and neutral words. Thus, despite major age-related differences in the initial item pool of 200 words, the selection process was successful in selecting age-equivalent words in terms of valence. This was quite important for the main question of this dissertation project whether older adults favor positive over negative information in memory and more so than young adults.

Unfortunately, the other rating dimensions revealed age-related differences in the selected item pool of positive, negative, and neutral words. On the one hand, young and older adults' ratings of the total item pool differed in self-relevance and age-relevance. Interestingly, young participants rated all words as somewhat more typical for themselves (i.e., self-relevance) than older participants did, whereas older participants rated all words as somewhat more typical for older people (i.e., age-relevance) than young participants did. This contradictory pattern for self-relevance and age-relevance is somewhat surprising. Nevertheless, it demonstrates that young- and older-stereotypes are not similar to the beliefs of young and older adults about their own personality. On the other hand, young and older adults' ratings for positive, negative, and neutral words differed in arousal and control. In contrast to older adults, young adults rated negative words as involving more arousal and less control and positive words as involving less arousal and more control. Young and older adults rated neutral words the same. This finding could indicate that older adults were less able to differentiate words on arousal and control.

In sum, the selection process was successful in realizing a final item pool of 30 positive, 30 negative, and 30 neutral words. The sets of positive, negative, and neutral words were matched on word frequency, word length, imagery, and age-relevance. Thus, with regards to memory-relevant characteristics, the selection process resulted in comparable sets of words. Moreover, young and older adults did not differ in their valence ratings for these word sets.