European Journal of Personality
Eur. J. Pers. **23**: 85–106 (2009)
Published online 26 January 2009 in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/per.705

How to Find What's in a Name: Scrutinizing the Optimality of Five Scoring Algorithms for the Name-letter Task

ETIENNE P. LeBEL* and BERTRAM GAWRONSKI
The University of Western Ontario, Canada

Abstract

Although the name-letter task (NLT) has become an increasingly popular technique to measure implicit self-esteem (ISE), researchers have relied on different algorithms to compute NLT scores and the psychometric properties of these differently computed scores have never been thoroughly investigated. Based on 18 independent samples, including 2690 participants, the current research examined the optimality of five scoring algorithms based on the following criteria: reliability; variability in reliability estimates across samples; types of systematic error variance controlled for; systematic production of outliers and shape of the distribution of scores. Overall, an ipsatized version of the original algorithm exhibited the most optimal psychometric properties, which is recommended for future research using the NLT. Copyright © 2009 John Wiley & Sons, Ltd.

Key words: name-letter task; initial preference task; implicit self-esteem; reliability; validity

INTRODUCTION

It has been over 20 years since Nuttin (1985, 1987) discovered the name-letter effect, the phenomenon whereby individuals prefer the letters of their name over non-name letters. Nuttin's original research was primarily concerned with determining the ubiquity of the name-letter effect in various European languages and understanding the mechanism underlying the effect, namely the affective consequences of 'owning' certain letters of the alphabet. A decade later, Greenwald and Banaji's (1995) treaty on implicit social cognition, which proclaimed implicit self-esteem (ISE) as an effect of the self-attitude on objects associated with the self, set the grounds for studies using the name-letter task (NLT) as a measure of ISE (Kitayama & Karasawa, 1997).

Since then, researchers have increasingly relied on the NLT to measure ISE, as evidence has accumulated showing the usefulness of the construct in understanding many facets of

*Correspondence to: Etienne P. LeBel, Department of Psychology, The University of Western Ontario, Social Science Centre, London, Ontario N6A 5C2 Canada. E-mail: elebel@uwo.ca

Received 7 October 2008
Revised 27 November 2008
Accepted 1 December 2008

human psychology. For example, using the NLT to measure ISE has helped shed light on depression (De Raedt, Schacht, Franck, & De Houwer, 2006; Franck, De Raedt, & De Houwer, 2007), physical health (Shimizu & Pelham, 2004), social acceptance (Baccus, Baldwin, & Packer, 2004), unrealistic optimism (Bosson, Brown, Zeigler-Hill, & Swann, 2003), feedback sensitivity (Dijksterhuis, 2004), self-regulation (Jones, Pelham, & Mirenberg, 2002) and defensiveness (Schröder-Abé, Rudolph, Wiesner, & Schütz, 2007).

Despite the successful use of the NLT in these studies, however, there are at least two problems that are serious enough to challenge the informational value of the reported findings. First, researchers have used different scoring algorithms to compute name-letter scores, a suboptimal situation that undermines the comparability of results across studies, and thereby the development of a cumulative knowledge base. A second issue, which is directly related to the first one, is that it is not known which algorithm is the most optimal one, because researchers rarely report reliability estimates for published findings using differently computed NLT scores and a systematic investigation comparing the psychometric properties of the different algorithms has never been conducted.

To address these issues, the main goal of the current study was to scrutinize the psychometric properties of five NLT scoring algorithms that have been used in the literature to determine the most optimal scoring strategy. The optimality of the scoring algorithms was evaluated in terms of three primary and two ancillary criteria, with greater weight given to the primary criteria. The primary criteria in assessing the NLT scoring algorithms were (a) reliability, (b) variability in reliability estimates across samples and (c) types of systematic error variance controlled for. The ancillary criteria were (d) systematic production of outliers and (e) shape of the distribution of scores. The distinction between these criteria was made to highlight the fact that the primary criteria (i.e. reliability, consistency, error) have a somewhat more solid basis compared to the 'softer' ancillary criteria (i.e. outliers, distribution). As secondary goals, we also aimed to obtain an overall estimate of the typical reliability of NLT scores and to examine correlations between the differently computed NLT scores and a common measure of explicit self-esteem (Rosenberg, 1965). Based on these analyses, we provide recommendations regarding the most optimal scoring algorithm for the NLT and draw preliminary conclusions about the typical reliability of NLT scores.

CURRENT ISSUES WITH THE NLT

As mentioned above, there are currently two major issues with the use of the NLT as a measure of ISE. First, researchers have used different scoring algorithms to compute name-letter scores in published papers. This situation is suboptimal for several reasons. First, using different algorithms undermines the comparability of research findings across studies. Hence, it remains unclear whether a finding reported in one study using a particular scoring algorithm is comparable to another finding using a different scoring algorithm, or if a published finding holds only for the particular algorithm used in the study (Bosson, Lakey, Campbell, Zeigler-Hill, Jordan, & Kernis, 2008). Second, inconsistent use of algorithms across studies can undermine the development of a cumulative knowledge base involving the construct of ISE as measured with the NLT. Specifically, it is possible that different algorithms yield different patterns of results, which affects accurate theorizing about the constructs involved (e.g. Gawronski & Bodenhausen, 2006). Third, the use of different algorithms ignores the fact that certain algorithms may be suboptimal due to

DOI: 10.1002/per

questionable psychometric properties. For example, some algorithms do not control for particular kinds of systematic error variance (e.g. individual differences in baseline response tendencies), whereas other algorithms may systematically produce outliers, which have the potential to distort the research findings obtained with these algorithms. Needless to say, these issues create a suboptimal situation for accurate theorizing about ISE.

A second major issue with the use of the NLT, which is directly related to the first one, is that it is currently unknown which of the five algorithms is the most optimal. This is the case because reliability estimates of findings using the NLT are rarely reported (e.g. Baccus et al., 2004; De Raedt et al., 2006; Dijksterhuis, 2004; Franck et al., 2007; Jones et al., 2002; Pelham, Koole, Hardin, Hetts, Seah, & DeHart, 2005; Riketta, 2005; Shimizu & Pelham, 2004; Verplanken, Friborg, Wang, Trafimow, & Woolf, 2007). Moreover, even if reliability estimates would be consistently reported, it can be argued that the decision of which algorithm is the most optimal should be based on more than just estimates of reliability (e.g. systematic error variance; shape of distribution; outliers). A systematic investigation of the psychometric properties of the different algorithms would certainly be helpful in this regard. The current contribution aims at filling this gap.

THE SCORING ALGORITHMS

Our review of the literature revealed five different algorithms, which we examined in the current investigation. In line with findings showing that name-letter effects tend to be stronger for initials compared to other letters of a person's name (e.g. Greenwald & Banaji, 1995), all algorithms entail computing the mean scores of preferences for the first and second initials. However, the five algorithms differ in terms of algebraic procedures that are completed before the two preference scores are combined into a single index of ISE.

1. *Baseline-corrected algorithm (B-algorithm)*: The most widely used algorithm, as popularized by Kitayama and Karasawa (1997), involves first calculating normative letter baselines by averaging the letter ratings for individuals whose initials do not include the letter. Then, the respective letter baselines are subtracted from the first and last name initial ratings. Algebraically, $\text{initials}_{\text{own}} - \text{initials}_{\text{baseline}}$. This algebraic procedure systematically controls for baseline differences in the attractiveness of different letters, which may result from differential exposure (Bornstein, 1989) or visual features of a given letter (Duckworth, Bargh, Garcia, & Chaiken, 2002).
2. *Self-corrected algorithm (S-algorithm)*: This algorithm involves the computation of a difference score between a participant's initial ratings and his or her mean ratings of the remaining letters of the alphabet (e.g. Schröder-Abé et al., 2007). Algebraically, $\text{initials}_{\text{own}} - \text{mean}_{\text{non-initials}}$. This algebraic procedure systematically controls for individual differences in baseline response tendencies in letter ratings, which may be rooted in various factors, including acquiescence (Jackson & Messick, 1958), individual differences in positive or negative affect (Watson, 1988) or transient mood states (Schwarz, 1990).
3. *Double-correction algorithm (D-algorithm)*: This strategy resembles a combination of the first two algorithms by controlling for both baseline differences in the attractiveness of different letters as well as individual differences in baseline response tendencies (e.g. Gawronski, Bodenhausen, & Becker, 2007). In a first step, first and last name initials are baseline corrected as in the B-algorithm. These scores are then

DOI: 10.1002/per

divided by each participant's mean rating of all letters of the alphabet. Algebraically, $(\text{initials}_{\text{own}} - \text{initials}_{\text{baseline}})/\text{mean}_{\text{allLetters}}$. This procedure controls for both baseline letter attractiveness and baseline response tendencies.

4. *Ipsatized double-correction algorithm (I-algorithm)*: This algorithm involves ipsatizing letter ratings to achieve a double-correction (e.g. Baccus et al., 2004). First, the mean rating of all non-initial letters is subtracted from each letter rating (analogous to the S-algorithm). Then, similar to how baselines are calculated in the B-algorithm, normative letter baselines are computed by averaging the ipsatized letter ratings for individuals whose initials do not include the letter. Finally, a difference score is computed between the ipsatized initial ratings and the respective ipsatized baselines. Algebraically, $(\text{initials}_{\text{own}} - \text{mean}_{\text{non-initials}}) - \text{initials}_{\text{ipsatizedBaseline}}$. As with the D-algorithm, this algorithm controls for both baseline levels of attractiveness of the different letters as well as individual differences in baseline response tendencies.
5. *Z-transformed double-correction algorithm (Z-algorithm)*: This algorithm involves first z-transforming each participant's letter ratings (each rating becomes a normal deviate from that person's mean level) and then calculating normative baselines on these z-transformed scores (as in the B-algorithm). Then, z-transformed letter baselines are subtracted from first and last name initial ratings (e.g. De Raedt et al., 2006). Algebraically, $\text{Zinitials}_{\text{own}} - \text{Zinitials}_{\text{baseline}}$. As with the D- and the I-algorithm, this algorithm systematically controls for both baseline differences in the attractiveness of different letters as well as individual differences in baseline response tendencies.

As should be apparent from their description, the five algorithms are similar in logic but differ with respect to which source of systematic error variance is controlled. For example, the B-algorithm controls for the baseline attractiveness of each letter, but does not control for individual differences in response tendencies of liking all letters more or less. In contrast, the S-algorithm controls for individual response tendencies, but does not control for the differences in the baseline attractiveness of each letter. The D-, I- and Z-algorithms control for both of these influences, although they achieve these corrections via slightly different calculus.

EVALUATION CRITERIA

Although all of these algorithms seem defensible on logical grounds, which of them is the most optimal algorithm in terms of its psychometric properties is an empirical question that needs to be investigated as such. The current investigation aimed to find the most optimal algorithm using five criteria, involving three primary and two ancillary criteria. Primary criteria included (a) reliability, (b) variability in reliability estimates across samples and (c) types of systematic error variance controlled for. Ancillary criteria involved (d) production of outliers and (e) shape of the distribution of scores.

These psychometric criteria were selected based on the reasoning that these characteristics reflect desirable or undesirable features from a psychometric or statistical perspective. First, given the importance of reliability for accurate measurement and replicability of findings (Nunnally, 1982), it is straightforward that algorithms yielding higher reliability estimates are more desirable (*reliability criterion*). Even though the reliability of implicit measures is often taken for granted in experimental research, one of the most important questions for individual difference research concerns the internal

DOI: 10.1002/per

consistency of these measures (Gawronski, 2009). Given that a major portion of research on ISE employs an individual difference design, reliability seems essential when it comes to evaluating the optimality of the five NLT algorithms. Second, algorithms that show more variability in reliability estimates across samples are less desirable than algorithms that yield more stable cross-sample reliability estimates (*consistency criterion*). This criterion is based on the quest for replicability, which requires comparable reliabilities of NLT scores to begin with. Third, it is clear that algorithms yielding NLT scores that tap distinctly into ISE are more desirable than algorithms in which variance due to ISE is confounded with other factors, for instance baseline levels of letter attractiveness and/or rating tendencies to prefer all letters more or less (*error criterion*). The latter confounding can be particularly troublesome, as individual differences in positive or negative affect (Watson, 1988) or transient mood states (Schwarz, 1990) may systematically influence baseline response tendencies in letter ratings, thereby undermining unambiguous interpretations of research findings in terms of ISE. Fourth, algorithms that, due to their underlying calculations, are more likely to produce extreme values are clearly less desirable than algorithms that are less likely to produce outliers (*outlier criterion*). Needless to say, outliers (retained or excluded) have the potential to systematically distort the findings revealed by a given measure, which can undermine accurate theorizing about ISE. Finally, from a statistical perspective, scores that are approximately normally distributed are more desirable than distributions of scores that are skewed or plagued with kurtosis (*distribution criterion*). This is the case because many statistical procedures (e.g. multiple regression) presuppose normal distributions of measurement scores, which if not met, can distort statistical estimates produced by these procedures (e.g. Vasu & Elmore, 1975).

METHOD

Participants and samples

Eighteen independent samples, with a total of 2690 distinct participants were examined in the current investigation. Table 1 presents demographic information for the samples.[1] Studies were conducted either with college students or internet samples. Overall, the average age of the total sample was slightly higher ($M = 22.8$) than the typical college sample because of the higher average age in the internet samples ($M_{\text{internet}} = 30.3$). All participants were included in all analyses except if (a) they had missing letter ratings or missing initial information (average of 6.1% per sample) or (b) they had the same rating for all letters of the alphabet (average of 4.4% per sample).[2] The vast majority of individuals participated for either course credit (college samples) or without compensation (internet samples). In all samples except the internet samples, participants were required to make a series of intuitive liking judgments of each letter of the alphabet, which were presented in a fixed random order. For the three internet samples, all letters of the alphabet were presented at once and in alphabetical order. All studies were computer based.

[1]As indicated in Table 1, samples 1, 2, 16 and 17 included experimental manipulations before NLT measurement. We found it justified to nonetheless include these samples on the grounds that no significant differences emerged across experimental conditions for NLT scores in each of these samples. Samples 14 and 15 were drawn from Tracy, Cheng, Robins, and Trzesniewski (in press).

[2]Having identical ratings for all letters of the alphabet was interpreted as non-compliance. In addition, identical ratings for all letters undermine a proper calculation of the Z-algorithm, as one cannot divide by an *SD* of 0.

DOI: 10.1002/per

Table 1. Demographic information for the 18 samples

					Age		Sex		Race			
Sample #	Sample type	Variant	Scale	*N*	Mean	*SD*	% female	% male	% European/White	% Black	% Asian	% Other
1[m]	College	Liking	1–5	91	18.5	1.0	67.7	32.3	79.6	2.2	11.8	6.4
2[m]	College	Liking	1–5	81	19.4	4.1	81.5	18.5	54.3	0.0	33.3	12.3
3	College	Liking	1–5	76	18.9	1.3	39.2	60.8	64.6	2.5	21.5	11.4
4	College	Aesthetic	1–7	79	22.9	5.4	76.3	23.8	45.0	5.0	37.5	12.5
5	College	Liking	1–5	78	18.8	3.0	72.5	27.5	56.3	1.3	30.0	12.5
6	College	Liking	1–5	80	19.0	2.3	55.6	44.4	65.4	0.0	29.6	4.9
7	College	Liking	1–5	78	18.5	0.93	68.8	31.3	70.0	2.5	17.5	10.0
8	College	Liking	1–5	81	18.9	1.3	50.0	48.8	63.1	0.0	13.1	23.8
9	College	Liking	1–5	119	21.8	6.3	73.6	26.4	32.2	1.7	52.1	14.0
10	College	Aesthetic	1–5	91	22.9	6.9	82.4	17.6	45.1	4.4	38.5	12.1
11	College	Liking	1–5	71	18.8	2.8	70.9	22.8	62.0	0.0	22.8	15.2
12	College	Liking	0–8	728	18.8	3.3	69.6	29.9	74.7	2.6	17.1	5.6
13	Internet	Liking	1–6	326	28.1	11.4	67.2	31.8	76.7	4.4	4.2	14.7
14	College	Liking	1–7	50	19.2	7.4	61.2	38.8	16.9	8.5	49.3	19.7
15	College	Aesthetic	1–7	43	19.4	1.5	56.1	43.9	15.2	4.5	47.0	19.7
16[m]	College	Liking	1–5	85	18.6	1.1	68.6	31.4	70.1	3.4	16.1	10.3
17[m]	Internet	Liking	1–6	260	31.3	11.1	58.5	40.5	74.2	10.0	5.4	10.4
18	Internet	Liking	1–6	273	31.5	13.9	63.7	35.0	67.0	9.1	7.3	16.6
Overall				2690	22.8	6.1	66.6	32.8	65.8	4.1	18.5	11.2

Note: [m] indicates pre-NLT manipulation. In line with the United States Census Bureau and the Federal Office of Management and Budget, persons having origins in any part of Asia or the India subcontinent were considered Asian.

DOI: 10.1002/per

Procedure

The main procedure involved (a) computing name-letter scores using the five different scoring algorithms for each sample individually,[3] (b) computing, for each sample, internal consistency estimates using the five algorithms, overall effect sizes, first versus second name initial preferences effect sizes, and correlations to the Rosenberg (1965) self-esteem scale (RSES) and (c) computing sample-weighted averages of the relevant parameter estimates.

In service of the investigation's main goal, we computed sample-weighted averages of the reliability estimates of the NLT scores using the different algorithms, giving larger weights to larger samples. We also examined sample-weighted effect sizes using Cohen's (1988) *d* for the overall name-letter scores, and separately for first versus last name initial preference scores, again using the five different scoring algorithms. To further validate these sample-weighted parameter estimates, we also computed all relevant statistics on the combined sample ($N = 2690$) composed of all 18 samples (applying the same outlier exclusion criterion). Finally, we performed various analyses to gauge whether type of sample, variant of the NLT task (liking vs. aesthetic letter judgments), scale format, and participant sex influenced the reliability estimates of the NLT scores.

RESULTS

Preliminary analyses

Preliminary analyses of the NLT scores using the five algorithms revealed the existence of outliers that significantly distorted reliability estimates. For example, in sample 9, a Cronbach's α for the D-algorithm of $\alpha = .70$ decreased to $\alpha = .48$ after the exclusion of a participant with first and last initial preference scores of 7 *SD* below the mean. After careful scrutiny of the scatterplots and histograms of first and second initial preference scores, it was determined that a cutoff value of 4 *SD* was the most inclusive cutoff value that successfully excluded reliability-distorting extreme values. This procedure led to the exclusion of a total of 11 cases for analyses using individual samples and a total of 13 cases for analyses using the combined sample.

Primary criteria

Reliability criterion

Table 2 presents Cronbach's α and split-half correlation reliability estimates for the five scoring algorithms.[4] As can be seen, the S- and I-algorithm had the highest reliability (average Cronbach's α of $\alpha_{avg} = .48$ and $\alpha_{avg} = .47$, respectively), followed by the B-

[3]To ensure comparability across samples in terms of overall means and standard deviations, all letter ratings were converted to a 1–5 scale using the following linear transformation: $\text{convertedScore} = [[\frac{(originalScore-1)}{(highestAnchorOriginal-1)}] \times 4] + 1$ (except for analyses probing effects of scale where raw data were used).

[4]Because negative Cronbach's α values are not interpretable (due to the fact that the Spearman–Brown prophecy formula was not designed to handle negative average inter-item correlations), the negative values were replaced with zeroes and included in the calculation of the sample-weighted average reliability estimates. This course of action was taken because simply ignoring the uninterpretable negative values would over-inflate the reliability estimates for the algorithms with negative values. Note, however, that the interpretable negative split-half correlations were included.

DOI: 10.1002/per

Table 2. Internal consistency estimates of name-letter scores using the five scoring algorithms across the 18 samples

				Cronbach's α					Split-half correlations				
Sample #	Sample type	Variant	N	α_B	α_S	α_D	α_I	α_Z	r_B	r_S	r_D	r_I	r_Z
1[m]	College	Liking	91	.29	.42	.29	.38	.24	.17	.27	.17	.24	.14
2[m]	College	Liking	81	.60	.73	.58	.62	.52	.43	.57	.41	.45	.35
3	College	Liking	76	.31	.47	.17	.45	.17	.19	.31	.10	.30	.09
4	College	Aesthetic	79	.40	.33	.37	.39	.30	.25	.20	.23	.25	.18
5	College	Liking	77	.54	.70	.53	.65	.62	.38	.55	.36	.49	.45
6	College	Liking	80	.23	.44	.21	.38	.36	.13	.29	.12	.23	.22
7	College	Liking	78	.37	.31	.24	.31	.00[u]	.24	.19	.14	.19	−.06
8	College	Liking	80	.08	.39	.14	.33	.10	.04	.25	.08	.20	.05
9	College	Liking	118	.49	.55	.48	.57	.46	.33	.38	.32	.40	.30
10	College	Aesthetic	91	.00[u]	.29	.00[u]	.20	.00[u]	−.06	.17	−.06	.11	−.03
11	College	Liking	70	.18	.46	.12	.33	.00[u]	.13	.33	.08	.22	−.03
12	College	Liking	725	.47	.51	.42	.50	.32	.32	.35	.27	.34	.19
13	Internet	Liking	324	.37	.42	.31	.42	.22	.23	.27	.19	.27	.12
14	College	Liking	49	.64	.41	.60	.43	.30	.47	.26	.43	.28	.18
15	College	Aesthetic	43	.49	.65	.52	.56	.54	.32	.48	.35	.39	.37
16[m]	College	Liking	85	.49	.51	.51	.59	.45	.33	.37	.35	.42	.29
17[m]	Internet	Liking	260	.51	.59	.49	.58	.42	.35	.42	.33	.41	.27
18	Internet	Liking	272	.43	.43	.39	.46	.28	.28	.27	.24	.30	.16
Weighted reliability estimates				.43(.15)	.48(.13)	.38(.16)	.47(.13)	.33(.15)	.27(.14)	.33(.12)	.24(.14)	.32(.10)	.18(.14)
Combined sample ($N = 2677$)				.42	.51	.38	.50	.33	.27	.34	.24	.33	.20

Note: α_B, B-algorithm; α_S, S-algorithm; α_D, D-algorithm; α_I, I-algorithm; α_Z, Z-algorithm; [m] indicates pre-NLT manipulation; [u] indicates that Cronbach's α was an undefined negative value. Values in parentheses represent the standard deviations of reliability estimates across the samples.

DOI: 10.1002/per

algorithm ($\alpha_{avg} = .43$), the D-algorithm ($\alpha_{avg} = .38$) and the Z-algorithm ($\alpha_{avg} = .33$). To examine whether the reliability estimates were significantly different from each other, we executed tests on the split-half correlations among the five algorithms from the combined sample (see Table 2). For this purpose, we used a quadratic form asymptotic χ^2 test (Steiger, 1980a, 1980b), involving a correlational pattern hypothesis whereby correlations between first and second initial preference scores for each algorithm were set equal. This analysis revealed that the split-half correlations among the five algorithms were significantly different from each other, $\chi^2(5) = 1029.4$, $p < .001$. *Post-hoc* analyses revealed that the split-half correlation for the I-algorithm ($r_I = .33$) was significantly larger than the B-algorithm ($r_B = .27$), $\chi^2(2) = 495.4$, $p < .001$. Correspondingly, the S- and I-algorithm had significantly higher reliability estimates than the D- and Z-algorithms (all χ^2[2]s > 410.0, all ps $< .001$).

As can be seen in Table 2, and substantiating the validity of our analysis, NLT scores computed on the 18 samples combined yielded reliability estimates that closely converged with the sample-weighted averages.[5] Another noteworthy aspect is that three of the five algorithms produced undefined (i.e. negative) Cronbach's α values in at least one of the subsamples (i.e. the B-, D- and Z-algorithms). These cases indicate that preference scores of the first initial were negatively correlated with preference scores of the second initial (see Table 2). Such negative correlations corrupt the required aggregation of first and second initial scores in a single index, as a negative correlation indicates that the two preference scores may tap conceptually distinct constructs.

Consistency criterion

As is evident from Table 2, there was notable variability in reliability estimates across samples and scoring algorithms (e.g. B-algorithm estimates ranged from an undefined $-.13$ to .64). To investigate variations in reliability across samples, we calculated the standard deviations of reliability estimates for the five algorithms (see Table 2). Standard deviations for the D-, Z- and B-algorithms ($SD_D = 0.16$, $SD_Z = 0.15$ and $SD_B = 0.15$, respectively) showed higher variability than the S- and I-algorithms ($SD_S = 0.13$ and $SD_I = 0.13$, respectively). To test whether these standard deviations were significantly different from each other, we tested differences in standard deviations on the split-half correlations to be consistent with our reliability criterion analyses. A dependent standard deviation test (Krause & Metzler, 1984), using sample as unit of analysis, revealed that the I-algorithm ($SD = 0.10$) showed significantly less variability as compared to the Z-algorithm ($SD = 0.14$), $t(16) = -2.54$, $p = .02$ and as compared to the D-algorithm ($SD = 0.14$), $t(16) = -2.32$, $p = .04$. All other comparisons were non-significant (all ps $> .05$).[6]

Ancillary criteria

Outlier criterion

Our analyses revealed that certain algorithms were more likely to produce outliers as compared to other algorithms. The D-, B- and Z-algorithms yielded 8, 4 and 3 outliers,

[5]To investigate a speculation that NLT scores computed using more 'stable' population letter baselines might be more reliable, we computed global normative baselines for the B-, D-, I- and Z-algorithms on the combined sample. These analyses revealed no overall difference between reliability estimates of NLT scores computed using the global ('population') baselines compared to the standard method of using sample-specific letter baselines.
[6]Analyses based exclusively on samples adopting the liking variant of the NLT revealed the same pattern of results.

DOI: 10.1002/per

respectively, whereas the I- and S-algorithm produced no outliers (total number of outliers was greater than 11 because certain cases were outliers on more than one algorithm). Examining a more restrictive cutoff value of 3 *SD* away from the mean revealed a similar pattern whereby the D-, B- and Z-algorithms yielded 24 outliers each, whereas the I- and S-algorithms produced 7 and 10 outliers, respectively.

Distribution criterion

Table 3 presents skewness and kurtosis statistics computed on the 18 samples (statistical significance determined by forming 95% confidence intervals around the statistics and checking whether the interval included zero). In terms of skewness, although most distributions were significantly negatively skewed across all five algorithms, the I-algorithm exhibited the most non-skewed distributions (i.e. nine non-skewed samples), with the S- and Z-algorithms exhibiting seven non-skewed distributions. In contrast, the B-algorithm had only three non-skewed distributions whereas the D-algorithm had skewed distributions in all samples. Concerning kurtosis, most samples did not suffer from significant kurtosis. The I-, S- and B-algorithms were the least plagued with kurtosis with only one sample each for the I- and S-algorithms, and two samples for the B-algorithm. In contrast, the Z- and D-algorithms had six and seven samples plagued with significant kurtosis, respectively. Consistent with these analyses, skewness and kurtosis statistics computed on the combined sample revealed that the I-, S- and Z-algorithms were less negatively skewed than the B- and D-algorithms and that the I-, S- and B-algorithms were plagued with less kurtosis than the D- and Z-algorithms. Also consistent with these analyses, one-sample Kolmogorov–Smirnov tests on the combined sample showed that the I-, S- and Z-algorithms ($Z = 2.43$, $Z = 3.14$ and $Z = 3.29$, respectively) deviated less from normality compared to the D- and B-algorithms ($Z = 4.88$ and $Z = 5.42$, respectively).

Secondary analyses

Effect sizes

Table 4 presents sample-specific and overall sample-weighted Cohen's *d* estimates of effect sizes (see Table 5 for means and standard deviations). Overall, effect sizes were large across all five scoring algorithms, $d_B = 1.24$, $d_S = 1.39$, $d_D = 1.19$, $d_I = 1.17$ and $d_Z = 1.31$. More interestingly, the sample-weighted first name initial effect sizes ($d_B = 1.27$, $d_S = 1.45$, $d_D = 1.20$, $d_I = 1.18$ and $d_Z = 1.25$) were considerably greater than the second initial effect sizes ($d_B = .77$, $d_S = .87$, $d_D = .72$, $d_I = .76$ and $d_Z = .79$).

Correlations to Rosenberg scale

Table 6 presents sample-specific correlations between NLT and Rosenberg scores for the five algorithms. As is evident from the table, the B- and D-algorithms yielded sample-weighted NLT–RSES correlations that were approximately twice as large as the Z-, I- and S-algorithms. Computed on the combined sample, a quadratic form asymptotic χ^2 test (Steiger, 1980a, 1980b) revealed that the correlations to RSES scores were significantly different from each other, $\chi^2(5) = 119.5$, $p < .001$. *Post-hoc* analyses showed that the RSES correlations of the B- and D-algorithms were significantly greater than the correlations of the I-, S- and Z-algorithms, all χ^2s > 51.0, all *p*s $< .001$. Although the obtained correlations seem small, the results are consistent with a recent meta-analysis by Krizan and Suls (2008) who found an overall NLT–RSES correlation of $r = .12$ across 19 independent samples (which likely used different NLT scoring algorithms). In addition, our NLT data converge

DOI: 10.1002/per

Table 3. Skewness and kurtosis statistics of name-letter scores for the five scoring algorithms across the 18 samples

Sample #	Sample type	Variant	N	Skewness					Kurtosis				
				B-algorithm	S-algorithm	D-algorithm	I-algorithm	Z-algorithm	B-algorithm	S-algorithm	D-algorithm	I-algorithm	Z-algorithm
1[m]	College	Liking	91	−1.23*	−.80*	−1.31*	−.65*	−.56*	1.81*	1.05*	2.69*	.58	.88
2[m]	College	Liking	81	−1.16*	−1.01*	−1.22*	−1.03*	−1.05*	.90	.58	1.18*	1.10*	1.50*
3	College	Liking	76	−.90*	−.92*	−.91*	−.78*	−.73*	.04	.63	.32	.01	1.25*
4	College	Aesthetic	79	−.40	−.39	−.59*	−.30	−.32	−.65	−.71	−.21	−.81	−.77
5	College	Liking	77	−1.07*	−.98*	−.97*	−.69*	−.93*	.52	.10	.62	.09	.88
6	College	Liking	80	−.59*	−.66*	−.59*	−.50	−.66*	−.55	−.14	−.34	−.33	−.09
7	College	Liking	78	−.85*	−.84*	−.93*	−.65*	−.35	.10	.26	.46	.32	.78
8	College	Liking	80	−.67*	−.64*	−.65*	−.54*	−.71*	−.32	.01	−.24	−.33	.21
9	College	Liking	118	−.98*	−.90*	−.99*	−.65*	−.31	.71	.70	.89*	.03	1.17*
10	College	Aesthetic	91	−.89*	−.24	−.86*	−.38	−.55*	.46	−.06	.85	−.28	−.54
11	College	Liking	70	−.92*	−.45	−.84*	−.38	.06	−.11	−.42	−.08	−.57	1.22*
12	College	Liking	725	−.88*	−.48*	−.87*	−.42*	−.28*	.12	.12	.59*	.05	.94*
13	Internet	Liking	324	−.86*	−.15	−.98*	−.05	−.09	.45	−.28	1.39*	−.12	.86*
14	College	Liking	49	−.59	−.28	−1.02*	−.21	−.16	−.53	−.88	.77	−.57	−.32
15	College	Aesthetic	43	−.60	−.28	−.72*	−.26	.36	.01	.62	.71	.51	.46
16*	College	Liking	85	−1.07*	−.57*	−1.00*	−.54*	−.88*	1.10*	.12	1.70*	.32	.45
17*	Internet	Liking	260	−.82*	−.30*	−.85*	−.17	−.43*	.02	−.27	.32	−.22	.55
18	Internet	Liking	272	−.89*	−.06	−1.04*	−.11	−.29*	.28	−.46	.98*	−.48	.33
Weighted statistics				−.87	−.46	−.92	−.38	−.37	.25	−.01	.75	−.07	.67
Combined sample ($N = 2677$)				−.92*	−.41*	−.96*	−.36*	−.32*	.35*	−.12	.91*	−.14	.74*

Note: [m] indicates pre-NLT manipulation. Positive (negative) skewness values indicate that most scores fall on the low end (high end) of the scale. Positive (negative) kurtosis values describe pointed (flat) distributions.

*indicates that the respective scores are significantly different from zero ($p < .05$).

Eur. J. Pers. **23**: 85–106 (2009)
DOI: 10.1002/per

Table 4. Cohen's d effect sizes of name-letter scores (d), first initial preference score (d_{1st}) and second initial preference score (d_{2nd}) for the five scoring algorithms across the 18 samples

				B-algorithm			S-algorithm			D-algorithm			I-algorithm			Z-algorithm		
Sample #	Sample type	Variant	N	d	d_{1st}	d_{2nd}	d	d_{1st}	d_{2nd}	d	d_{1st}	d_{2nd}	d	d_{1st}	d_{2nd}	d	d_{1st}	d_{2nd}
1[m]	College	Liking	91	1.29	1.24	0.80	1.44	1.36	0.96	1.22	1.16	0.75	1.21	1.14	0.79	1.29	1.18	0.79
2[m]	College	Liking	81	1.06	0.99	0.80	1.26	1.36	0.89	1.03	0.97	0.77	1.03	0.95	0.81	1.15	1.00	0.89
3	College	Liking	76	1.47	1.60	0.81	1.52	1.60	0.92	1.45	1.52	0.76	1.31	1.39	0.79	1.36	1.21	0.79
4	College	Aesthetic	79	0.79	0.77	0.47	0.86	0.83	0.49	0.74	0.75	0.42	0.79	0.74	0.50	0.90	0.84	0.53
5	College	Liking	77	1.29	1.25	0.91	1.25	1.36	0.87	1.25	1.19	0.89	1.11	1.08	0.85	1.14	1.16	0.80
6	College	Liking	80	1.06	1.00	0.62	1.22	1.22	0.76	1.04	1.00	0.58	0.96	0.90	0.60	1.03	0.95	0.65
7	College	Liking	78	1.47	1.62	0.83	1.89	2.01	1.06	1.44	1.58	0.76	1.54	1.58	0.88	1.67	1.39	0.89
8	College	Liking	80	1.95	1.43	1.39	1.88	1.54	1.43	1.94	1.45	1.39	1.67	1.27	1.32	1.89	1.43	1.31
9	College	Liking	118	1.30	1.32	0.84	1.37	1.42	0.90	1.27	1.28	0.82	1.18	1.18	0.80	1.28	1.26	0.82
10	College	Aesthetic	91	1.19	1.00	0.63	1.20	1.09	0.73	1.15	0.97	0.61	1.01	0.92	0.58	1.17	1.01	0.61
11	College	Liking	70	1.69	2.20	0.87	1.82	2.25	1.02	1.65	2.14	0.82	1.54	1.81	0.86	1.67	2.07	0.82
12	College	Liking	725	1.40	1.51	0.85	1.57	1.71	0.95	1.35	1.43	0.80	1.35	1.44	0.84	1.51	1.47	0.89
13	Internet	Liking	324	1.15	1.21	0.67	1.33	1.44	0.77	1.07	1.12	0.60	1.08	1.12	0.65	1.25	1.23	0.69
14	College	Liking	49	0.77	0.82	0.52	1.14	1.09	0.72	0.65	0.71	0.41	0.92	0.86	0.60	0.96	0.90	0.57
15	College	Aesthetic	43	0.67	0.66	0.43	0.65	0.68	0.44	0.56	0.62	0.32	0.62	0.58	0.44	0.70	0.68	0.48
16[m]	College	Liking	85	1.45	1.66	0.81	1.55	1.90	0.86	1.39	1.59	0.78	1.29	1.46	0.78	1.46	1.54	0.85
17[m]	Internet	Liking	260	1.00	0.99	0.67	1.15	1.18	0.78	0.95	0.94	0.63	0.92	0.91	0.66	1.05	1.01	0.68
18	Internet	Liking	272	1.03	0.93	0.72	1.24	1.11	0.86	0.97	0.84	0.68	1.01	0.91	0.71	1.15	1.04	0.72
Weighed effect sizes				1.24	1.27	0.77	1.39	1.45	0.87	1.19	1.20	0.72	1.17	1.18	0.76	1.31	1.25	0.79
Combined sample ($N = 2677$)				1.28	1.21	0.79	1.35	1.38	0.86	1.24	1.24	0.77	1.18	1.17	0.72	1.15	1.13	0.76

Note: [m] indicates pre-NLT manipulation.

Eur. J. Pers. **23**: 85–106 (2009)
DOI: 10.1002/per

Table 5. Means and standard deviations of name-letter scores for the five scoring algorithms across the 18 samples

				Means (standard deviations)				
Sample #	Sample type	Variant	*N*	B-algorithm	S-algorithm	D-algorithm	I-algorithm	Z-algorithm
1[m]	College	Liking	91	1.06 (.82)	1.28 (.89)	0.33 (.27)	1.07 (.88)	0.86 (.66)
2[m]	College	Liking	81	0.96 (.90)	1.15 (.92)	0.29 (.28)	0.96 (.93)	0.82 (.71)
3	College	Liking	76	1.11 (.75)	1.24 (.82)	0.34 (.23)	1.11 (.84)	1.01 (.74)
4	College	Aesthetic	79	0.72 (.92)	0.78 (.91)	0.22 (.30)	0.72 (.92)	0.64 (.71)
5	College	Liking	77	1.00 (.77)	1.15 (.92)	0.30 (.24)	1.00 (.90)	0.83 (.73)
6	College	Liking	80	0.87 (.82)	1.07 (.88)	0.26 (.25)	0.87 (.91)	0.76 (.74)
7	College	Liking	78	1.05 (.72)	1.25 (.66)	0.31 (.22)	1.06 (.69)	0.97 (.58)
8	College	Liking	80	1.19 (.61)	1.37 (.73)	0.37 (.19)	1.20 (.71)	1.02 (.54)
9	College	Liking	118	1.02 (.78)	1.17 (.84)	0.31 (.24)	1.00 (.85)	0.86 (.68)
10	College	Aesthetic	91	0.91 (.76)	1.08 (.90)	0.28 (.25)	0.90 (.90)	0.78 (.67)
11	College	Liking	70	1.17 (.69)	1.37 (.75)	0.36 (.22)	1.17 (.76)	0.96 (.57)
12	College	Liking	725	1.01 (.72)	1.17 (.74)	0.31 (.23)	1.00 (.74)	1.06 (.70)
13	Internet	Liking	324	0.72 (.63)	0.88 (.66)	0.19 (.18)	0.71 (.66)	0.81 (.65)
14	College	Liking	49	1.02 (1.33)	1.15 (1.01)	0.19 (.29)	0.97 (1.06)	0.86 (.89)
15	College	Aesthetic	43	0.87 (1.33)	0.97 (1.48)	0.19 (.35)	0.88 (1.43)	0.74 (1.06)
16[m]	College	Liking	85	1.12 (.77)	1.32 (.85)	0.36 (.26)	1.12 (.87)	0.90 (.62)
17[m]	Internet	Liking	260	0.67 (.67)	0.84 (.73)	0.18 (.19)	0.67 (.72)	0.70 (.67)
18	Internet	Liking	272	0.69 (.66)	0.82 (.66)	0.18 (.19)	0.69 (.68)	0.77 (.67)
Weighted means				0.91 (.74)	1.07 (.78)	0.27 (.22)	0.91 (.78)	0.88 (.68)
Combined sample ($N = 2677$)				0.89 (.69)	1.06 (.78)	0.26 (.22)	0.91 (.73)	0.90 (.79)

Note: [m] indicates pre-NLT manipulation. Standard deviations are presented in parentheses.

Eur. J. Pers. **23**: 85–106 (2009)
DOI: 10.1002/per

Table 6. Correlations between NLT and RSES scores for the five scoring algorithms across the 18 samples

				Correlations				
Sample #	Sample type	Variant	N	r_B	r_S	r_D	r_I	r_Z
1[m]	College	Liking	91	.30	.33	.32	.34	.32
2[m]	College	Liking	81	.24	.18	.23	.16	.15
3	College	Liking	76	.03	−.02	.03	−.02	.02
4	College	Aesthetic	79	.08	.08	.08	.09	.03
5	College	Liking	77	.13	.01	.09	−.03	−.08
6	College	Liking	80	.06	.10	.04	.06	.12
7	College	Liking	78	.16	−.01	.12	.04	−.04
8	College	Liking	80	.05	.03	.05	.04	.04
9	College	Liking	118	.07	.05	.05	.01	.01
10[†]	College	Aesthetic	91	—	—	—	—	—
11	College	Liking	70	−.03	.02	−.05	−.04	.04
12	College	Liking	725	.18	.10	.15	.10	.11
13	Internet	Liking	324	.21	−.05	.18	−.02	.03
14	College	Liking	49	.20	.01	.15	.05	.19
15	College	Aesthetic	43	.27	.08	.30	.12	−.01
16[m]	College	Liking	85	.09	.15	.09	.09	.09
17[m]	Internet	Liking	260	.22	.07	.20	.08	.07
18	Internet	Liking	272	.29	.09	.27	.11	.13
Weighted correlations				.17 (.10)	.07 (.09)	.15 (.10)	.07 (.09)	.08 (.09)
Combined sample ($N = 2677$)				.10	.02	.09	.02	−.05

Note: r_B, B-algorithm; r_S, S-algorithm; r_D, D-algorithm; r_I, I-algorithm; r_Z, Z-algorithm; [m] indicates pre-NLT manipulation.
[†]indicates the sample did not include the Rosenberg scale and thus correlations could not be computed. Values in parentheses indicate the standard deviations of the mean correlations across samples.

DOI: 10.1002/per

to meta-analytic findings on the self-esteem IAT (Greenwald & Farnham, 2000), showing an error-corrected, mean population correlation of .13 (Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005a).

Miscellaneous secondary analyses

In terms of judgment type, liking judgments ($N = 2466$) exhibited marginally higher split-half correlations than aesthetic judgments ($N = 213$) for the B- and I-algorithms (B-algorithm: $r_{liking} = .28$ vs. $r_{aesthetic} = .16$, $z = 1.79$, $p = .07$; I-algorithm: $r_{liking} = .35$ vs. $r_{aesthetic} = .23$, $z = 1.78$, $p = .07$). There were no differences between the S-, D-, and Z-algorithms (all $ps > .12$). Consistent with this finding, three of the five negative Cronbach's α values came from aesthetic judgment data, suggesting that aesthetic judgments may be less reliable than liking judgments. However, these differences must be interpreted with caution given the relatively small sample of aesthetic judgments. In terms of sample type, split-half correlations computed on the combined college ($N = 1824$) and internet ($N = 856$) samples were not significantly different from each other across the five algorithms (all $ps > .50$). We also examined whether NLT scores from samples using different scales (e.g. 1–5 vs. 1–6 vs. 1–7 vs. 0–8) differed in reliability across the algorithms. This analysis did not reveal any significant differences (all $ps > .14$). Finally, there were also no sex differences in reliability estimates across the five algorithms (all $ps > .60$).

DISCUSSION

The present investigation aimed to examine the optimality of five scoring algorithms of the NLT in terms of (a) reliability, (b) variability in reliability estimates across samples, (c) types of systematic error variance controlled for, (d) systematic production of outliers and (e) shape of the distribution of scores. In addition to giving more weight to primary (reliability, consistency, error) versus ancillary (outlier, distribution) evaluation criteria, specific issues relevant to the different algorithms need to be considered in determining the most optimal scoring algorithm (see Table 7 for a performance summary of the five algorithms with respect to the five evaluation criteria).

Table 7. Performance summary of the five NLT algorithms in terms of the five evaluation criteria.

Criterion	B-algorithm	S-algorithm	D-algorithm	I-algorithm	Z-algorithm
Reliability[a]	$\alpha = .43/.42$ 1 undefined α	$\alpha = .48/.51$ 0 undefined α	$\alpha = .38/.38$ 1 undefined α	$\alpha = .47/.50$ 0 undefined α	$\alpha = .33/.33$ 3 undefined α
Consistency	$SD = 0.15$	$SD = 0.13$	$SD = 0.16$	$SD = 0.13$	$SD = 0.15$
Error controlled for	Baseline letter attractiveness only	Baseline response tendencies only	Both	Both	Both
Outlier	4 outliers	0 outliers	8 outliers	0 outliers	3 outliers
Distribution	$S = -.92$ $K = .35$	$S = -.41$ $K = -.12$	$S = -.96$ $K = .91$	$S = -.36$ $K = -.14$	$S = -.32$ $K = .74$

Note: [a] The first α-value reported in the respective cells reflects the weighted mean reliability of a given algorithm; the second α-value reflects the reliability of the same algorithm for the combined sample. *S* and *K* values refer to skewness and kurtosis statistics, for the combined sample respectively.

DOI: 10.1002/per

Performance summary

B-algorithm

The B-algorithm fared relatively well in terms of the reliability criterion, achieving the third highest reliability estimate and yielding only one negative Cronbach's α value. However, the algorithm performed less favourably in terms of the consistency criterion, showing relatively high standard deviations in reliability estimates among the five algorithms. In addition, the B-algorithm performed less well in terms of the distribution criterion, producing one of the most skewed deviation from normal distribution. With regard to the outlier criterion, the B-algorithm also appeared suboptimal, yielding four outliers according to our cutoff criterion. Finally, the algorithm is suboptimal with regard to the error criterion, as it does not control for systematic error from baseline response tendencies of judging all letters more or less favourably. This issue seems important because such baseline response tendencies may be the result of individual differences in positive and negative affect (Watson, 1988) or transient mood states (Schwarz, 1990). Such confoundings have the potential to distort theorizing about ISE, if effects that are produced by these variables are misattributed to ISE.

S-algorithm

The S-algorithm performed very well in terms of the outlier criterion, not yielding a single outlier. It also appeared first rate with regard to the reliability criterion, yielding the highest reliability estimates and not producing any negative Cronbach's α. The algorithm also fared reasonably well with respect to the consistency and distribution criteria. However, the S-algorithm was not well suited in terms of the error criterion. Specifically, the algorithm does not control for baseline levels of letter favourability. As such, scores produced by this algorithm do not distinctively reflect ISE, which undermines unambiguous interpretations of the scores revealed by this algorithm.

D-algorithm

The D-algorithm fares very well with respect to the error criterion, as it controls for both baseline letter attractiveness and baseline response tendencies. However, the D-algorithm performed suboptimally in terms of all other criteria. First, the D-algorithm produced the highest number of outliers, which also turned out to be the most extreme outliers.[7] The algorithm also fared poorly with regard to the reliability criterion, yielding the second lowest reliability estimate and generating one negative Cronbach's α. With respect to the consistency criterion, the algorithm also did not seem well suited, as it produced the highest variability in reliability estimates across samples. Finally, the D-algorithm appeared suboptimal in terms of the distribution criterion, yielding samples plagued with the highest skew and kurtosis.

I-algorithm

The I-algorithm appeared well suited in terms of the outlier criterion, producing not a single outlier. The algorithm also fared well in terms of the reliability criterion, yielding the second highest reliability estimate and not a single negative Cronbach's α. The I-algorithm

[7]In this context, it is important to note that the D-algorithm is the only algorithm which does not yield meaningful scores when letter ratings originate from a scale that includes a zero point. This is the case when, on a 0 to x (e.g. 5) scale, a participant frequently endorses the zero point on the scale. In such cases, initial preference scores divided by a number smaller than one can produce two very large D-scores, and thus extreme outliers.

DOI: 10.1002/per

was also optimal in terms of the consistency and distribution criteria, producing the highest level of consistency across reliability estimates and exhibiting the lowest levels of skewness and kurtosis among the algorithms. Finally, the algorithm is also well suited in terms of the error criterion, as it controls for both baseline levels of letter favourability and baseline response tendencies.

Z-algorithm

The Z-algorithm is optimal in terms of the error criterion, as it controls for both baseline letter attractiveness and baseline response tendencies. In terms of the distribution criteria, the Z-algorithm fared reasonably well in terms of skewness, but not so well in terms of kurtosis. In addition, the Z-algorithm appeared suboptimal in terms of the outlier criterion, yielding three outliers. The algorithm was also not well suited in terms of the reliability criterion, yielding the lowest reliability estimate among all algorithms and producing the highest number of negative Cronbach's α. Finally, the algorithm performed suboptimally in terms of the consistency criterion, showing relatively large variability in reliability estimates across samples.

Summary

Taken together, our results indicate that the I-algorithm exhibited the best performance in terms of its psychometric properties, especially with regards to the primary evaluation criteria. Compared to the other four algorithms, the I-algorithm showed the highest reliability estimates (in the range of $\alpha = .50$, together with the S-algorithm), showed relatively low variability in reliability estimates across samples, and importantly controlled for both types of systematic error in the NLT (i.e. baseline letter attractiveness and baseline response tendencies), thus allowing for unambiguous interpretations in terms of ISE rather than other sources of systematic variance (e.g. positive or negative affect; transient mood states). The I-algorithm also seemed to be the best suited algorithm concerning the ancillary criteria, given that the algorithm did not produce a single outlier and exhibited relatively low deviation from normality.

Overall reliability

A secondary question was to provide an estimate of the overall reliability that emerged from the $N = 2677$ combined sample. Using the I-algorithm as the most optimal one, we obtained an overall estimate of $\alpha = .50$. On the one hand, it can be argued that this score is clearly unsatisfactory, because according to classical test theory, a reliability of .50 indicates that 50% of the variability in NLT scores is random error. Hence, it seems clear that viewed in this light, the research findings involving the NLT must be interpreted with caution. Specifically, it seems important to realize that effects using the NLT may be more difficult to replicate given that the scores may be contaminated with a relatively high degree of random influences. On the other hand, however, one could argue that according to the Spearman–Brown prophecy formula, a $\alpha = .50$ translates into an inter-item correlation of $r = .34$, which is reasonably high given that there are only two items in the NLT and that the indirect nature of the measure allows more opportunity for noise to contaminate the scores as compared to direct measures (but see Krizan, 2008). From this viewpoint, one should judge the merit of an inter-item correlation of $r = .34$ against inter-item correlations of other scales. For example, the directly transparent 10-item Rosenberg scale typically shows Cronbach's α estimates in the range of .85, which translates into an average inter-item

DOI: 10.1002/per

correlation of $r = .36$. Thus, it follows that the NLT's inter-item correlation of $r = .34$ fares quite well viewed in this light. Granted, this kind of argument is not intended to suggest that low reliability is not a problem. Rather, it allows for the judgment of the NLT's psychometric properties in terms of more reasonable standards. Hence, although the typical reliability of the NLT is clearly inadequate in a strict sense, the inter-item correlation of $r = .34$ provides a reasonable justification to use the measure as a proxy of ISE, at least when evaluated in terms of reliability.

A noteworthy finding in the context of discussing overall reliability is that the aesthetic version of the NLT produced lower reliability estimates compared to the liking version (at least with respect to the I- and B-algorithms). Although the small sample size for the aesthetic version limits the possibility of drawing strong conclusions, this finding suggests that the specific judgment employed in the task may affect the measurement of the construct. Future research should seek to confirm this finding to provide a better understanding of the possibly different psychological mechanisms underlying the two NLT variants (for related findings, see Sakellaropuollo & Baldwin, 2007).

Another interesting result is the finding that effect sizes for first initials were considerably larger than second initials. In line with arguments offered by Hetts, Sakuma, and Pelham (1999), it may well be that preference scores for first name initials are more meaningful, or at least more affectively powerful, than second name initials (see also Kitayama & Karasawa, 1997). Although questionable from a psychometric perspective, it seems interesting whether using only first initial preference scores may provide a better proxy of ISE than an average of preference scores for first and second name initials. In line with this contention, Shimizu and Pelham (2004) used only participant's first initial preference scores, although they mentioned in a footnote that their findings held when using the average of both. Future research may provide deeper insights in this regard by systematically investigating the validity of first name and last name initials as measures of ISE.

Relations to explicit self-esteem

Another secondary question of this research was to examine correlations between the differently computed NLT scores and the RSES. Overall, the B- and D-algorithms produced the highest correlations between NLT and RSES scores, with the S-, I- and Z-algorithms revealing correlations that were significantly lower. However, in evaluating this finding, it is important to remember that the RSES is but one of many potential criterion measures to validate ISE measures. More seriously, it may not be the most suitable one, as it is not entirely clear how to interpret correlations between explicit and implicit measures in the absence of any additional information. In fact, previous research has shown that the relation between explicit and implicit measures varies as a function of multiple variables (for a review, see Hofmann, Gschwendner, Nosek, & Schmitt, 2005b), which makes it difficult to evaluate the validity of an implicit measure on the basis of its zero-order correlations with a corresponding explicit measure. On the one hand, low correlations can be interpreted as showing either that (a) the two measures tap two conceptually distinct constructs or (b) the implicit measure does not really capture the desired construct. On the other hand, high correlations can be interpreted as showing either that (a) the implicit measure does indeed tap the desired construct or (b) the implicit measure is contaminated by explicit processes. In line with these concerns, Bosson, Swann, and Pennebaker (2000) recommended that researchers should not use self-report self-esteem measures to validate ISE measures, but instead rely on indirect or nonconscious criterion measures.

DOI: 10.1002/per

Alternatively, another more suitable approach would be to take previously identified moderators of explicit–implicit relations into account by testing whether these variables also moderate the relation between NLE and RSES scores. Such investigations seem better suited to determine the validity of an implicit measure compared to a simple correlational approach (e.g. Fazio, Jackson, Dunton, & Williams, 1995; Payne, Cheng, Govorun, & Stewart, 2005).

Implications

The present results have important implications for researchers using the NLT. First, and most importantly, the I-algorithm is strongly recommended as the preferred scoring procedure to calculate NLT scores. This recommendation is based on the combination of favourable characteristics exhibited by the I-algorithm (i.e. highest reliability estimate, least variability in reliability estimates across samples, correction for both types of systematic error, no outliers, relatively low deviation from normality and no negative Cronbach's α). For the sake of building a cumulative body of knowledge involving ISE, it is essential to maintain consistency across studies in NLT scoring. The current paper makes a strong case for choosing the I-algorithm as the preferred algorithm of choice.

Second, the current analysis implies that, although the overall reliability of NLT scores was found to be relatively low from a strict psychometric perspective, this mediocre reliability is defensible on the basis that the NLT includes only two items. As outlined above, the reliability estimate obtained for the I-algorithm translates into an inter-item correlation that is comparable to those of traditional multi-item measures (e.g. the 10-item RSES, Rosenberg, 1965; or the 40-item Narcissistic Personality Inventory, Raskin & Terry, 1988). Granted, this does not imply that researchers should not strive to improve the psychometric properties of the NLT.

A further important recommendation is that researchers should consistently report sample-specific reliability estimates of NLT scores. As evident in the large fluctuations in reliability estimates across samples, it is important to interpret NLT results in the context of the reliability of the measure. A low NLT reliability implies that the finding may be difficult to replicate due to the large proportion of random factors impinging on the scores. In the spirit of this recommendation and to facilitate the computationally complex nature of the task, we have created SPSS compatible syntax of the five algorithms including reliability estimate calculations (available to download from http://publish.uwo.ca/∼elebel/NLT.html).

The current findings also highlight that no measure is perfect and that a multi-method approach is the most optimal path to proper theorizing. Although from a narrow-minded reliability perspective, one could argue that the NLT should be abandoned, it is important to remember that different measures are often characterized by different strengths and weaknesses. From this perspective, it is important to employ different measures to establish the validity of a given finding. In this light, the name-liking task (Gebauer, Riketta, Broemer, & Maio, 2008) and the self-esteem IAT (Greenwald & Farnham, 2000) may be useful tools to corroborate findings obtained with the NLT in the domain of ISE.

Limitations

Despite the clear superiority of the I-algorithm in terms of our primary and ancillary criteria, it seems appropriate to acknowledge a number of limitations in the present study. First, the current research did not include a strong validity criterion to corroborate our main conclusion.

DOI: 10.1002/per

Thus, it is unclear whether the I-algorithm is also superior in terms of criterion validity. In an ideal case, it would have been most compelling to provide evidence that the I-algorithm also afforded better predictive validity of a relevant ISE criterion measure. Alternatively, one could investigate whether previously identified moderators of explicit–implicit relations (see Hofmann et al., 2005b) influence the relation between I-scored NLT and RSES scores in a more consistent fashion. Unfortunately, the current data sets did not include the type of measures that are required for these analyses. Future research comparing the criterion validity of different algorithms or moderating effects in the prediction of self-reported self-esteem may help to further clarify the optimality of the five scoring algorithms.

A second limitation is that our samples were relatively homogeneous, in that the majority of our participants were unmarried Caucasian college students. As such, it is unclear whether our results generalize to other cultures or to samples with higher proportions of married individuals whose initials may have changed. Along the same lines, one could object that our results may not generalize to samples with different levels of education. However, in response to this concern, it is worth noting that our three internet samples ($N = 859$) were composed of participants with considerable variability in education levels (high school or less = 10.3%, some college = 34.6%, associate degree = 6.6%, bachelor's = 19.8%, some graduate training = 8.6% and graduate degree = 18.8%). Thus, this limitation may not be as severe as one might initially presume. One last limitation of the present work is that we examined preferences for initial letters rather than all of the name letters. However, given that it is now common practice to calculate NLT scores exclusively from initial letters, this strategy seems appropriate when it is evaluated from a pragmatic point of view.

CONCLUSION

The development of a cumulative and coherent body of knowledge hinges upon the sound measurement of constructs. As mentioned in the introductory paragraph, the examination of ISE shows great promise in increasing our understanding of many facets of human psychology, including depression (De Raedt et al., 2006; Franck et al., 2007), physical health (Shimizu & Pelham, 2004), social acceptance (Baccus et al., 2004), unrealistic optimism (Bosson et al., 2003), feedback sensitivity (Dijksterhuis, 2004), self-regulation (Jones et al., 2002) and defensiveness (Schröder-Abé et al., 2007). Given the promise of these findings, accurate measurement of ISE seems important to further our understanding of these phenomena. The present study aimed at closing a significant gap in this regard by providing evidence-based recommendations on the most optimal scoring of the NLT as a measure of ISE.

ACKNOWLEDGEMENTS

The present research was supported by doctoral fellowships by the Social Sciences and Humanities Research Council of Canada and the Ontario Graduate Scholarship Program to the first author, and by the Canada Research Chairs Program Grant 202555 and Social Sciences and Humanities Research Council of Canada Grant 410-2008-2247 to the second author. We thank Kali Trzesniewski for helpful discussions about this research and Jessica Tracy and Joey Cheng for providing two samples of data.

DOI: 10.1002/per

REFERENCES

Baccus, J. R., Baldwin, M. W., & Packer, D. J. (2004). Increasing implicit self-esteem through classical conditioning. *Psychological Science*, *15*, 498–502.

Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research 1968–1987. *Psychological Bulletin*, *106*, 265–289.

Bosson, J. K., Brown, R. P., Zeigler-Hill, V., & Swann, W. B., Jr. (2003). Self-enhancement tendencies among people with high explicit self-esteem: The moderating role of implicit self-esteem. *Self and Identity*, *2*, 169–187.

Bosson, J. K., Lakey, C. E., Campbell, W. K., Zeigler-Hill, V., Jordan, C. H., & Kernis, M. H. (2008). Untangling the links between narcissism and self-esteem: A theoretical and empirical review. *Social and Personality Psychology Compass*, *2*, 1415–1439.

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology*, *79*, 631–643.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

De Raedt, R., Schacht, R., Franck, E., & De Houwer, J. (2006). Self-esteem and depression revisited: Implicit positive self-esteem in depressed patients? *Behaviour Research and Therapy*, *44*, 1017–1028.

Dijksterhuis, A. (2004). I like myself but I don't know why: Enhancing implicit self-esteem by subliminal evaluative conditioning. *Journal of Personality and Social Psychology*, *86*, 345–355.

Duckworth, K. L., Bargh, J. A., Garcia, M., & Chaiken, S. (2002). The automatic evaluation of novel stimuli. *Psychological Science*, *13*, 513–519.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*, 1013–1027.

Franck, E., De Raedt, R., & De Houwer, J. (2007). Implicit but not explicit self-esteem predicts future depressive symptomatology. *Behaviour Research and Therapy, 45*, 2448–2455.

Gawronski, B. (2009). Ten frequently asked questions about implicit measures and their frequently supposed, but not entirely correct answers. *Canadian Psychology.* DOI: 10.1037/a0013848

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*, 692–731.

Gawronski, B., Bodenhausen, G. V., & Becker, A. (2007). I like it, because I like myself: Associative self-anchoring and post-decisional change of implicit attitudes. *Journal of Experimental Social Psychology*, *43*, 221–232.

Gebauer, J. E., Riketta, M., Broemer, P., & Maio, G. (2008). How much do you like your name? An implicit measure of global self-esteem. *Journal of Experimental Social Psychology*, *44*, 1346–1354.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4–27.

Greenwald, A. G., & Farnham, S. D. (2000). Using the implicit association test to measure self-esteem and self-concept. *Journal of Personality and Social Psychology*, *79*, 1022–1038.

Hetts, J. J., Sakuma, M., & Pelham, B. W. (1999). Two roads to positive regard: Implicit and explicit self-evaluation and culture. *Journal of Experimental Social Psychology*, *35*, 512–559.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005a). A meta-analysis on the correlation between the implicit association test and explicit self-report measure. *Personality and Social Psychology Bulletin*, *31*, 1369–1385.

Hofmann, W., Gschwendner, T., Nosek, B. A., & Schmitt, M. (2005b). What moderates explicit-implicit consistency? *European Review of Social Psychology*, *16*, 335–390.

Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin*, *55*, 243–252.

Jones, J. T., Pelham, B. W., & Mirenberg, M. (2002). Name letter preferences are not merely mere exposure: Implicit egotism as self-regulation. *Journal of Experimental Social Psychology*, *38*, 170–177.

Kitayama, S., & Karasawa, M. (1997). Implicit self-esteem in Japan: Name letters and birthday numbers. *Personality and Social Psychology Bulletin*, *23*, 736–742.

DOI: 10.1002/per

Krause, B., & Metzler, P. (1984). *Angewandte Statistik [Applied statistics]*. Berlin: VEB Deutscher Verlag der Wissenschaften.

Krizan, Z. (2008). What is implicit about implicit self-esteem? The case of the name-letter effect. *Journal of Research in Personality.* DOI: 10.1016/j.jrp.2008.07.002

Krizan, Z., & Suls, J. (2008). Are implicit and explicit measures of self-esteem related? A meta-analysis for the name-letter test. *Personality and Individual Differences*, *44*, 521–531.

Nunnally, J. C. (1982). Reliability of measurement. In H. E. Mitzel (Ed.), *Encyclopedia of educational research* (pp. 1589–1601). New York: Free Press.

Nuttin, M. J., Jr. (1985). Narcissism beyond Gestalt and awareness: The name letter effect. *European Journal of Social Psychology*, *64*, 723–739.

Nuttin, M. J., Jr. (1987). Affective consequences of mere ownership: The name letter effect in twelve European languages. *European Journal of Social Psychology*, *17*, 381–402.

Payne, B. K., Cheng, S. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*, 277–293.

Pelham, B. W., Koole, S. L., Hardin, C. D., Hetts, J. J., Seah, E., & DeHart, T. (2005). Gender moderates the relation between implicit and explicit self-esteem. *Journal of Experimental Social Psychology*, *41*, 84–89.

Raskin, R., & Terry, H. (1988). A principal-components analysis of the Narcissistic Personality Inventory and further evidence of its construct validity. *Journal of Personality and Social Psychology*, *54*, 890–902.

Riketta, M. (2005). Gender and socially desirable responding as moderators of the correlation between implicit and explicit self-esteem. *Current Research in Social Psychology*, *11*, 14–28.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Sakellaropuollo, M., & Baldwin, M. W. (2007). The hidden sides of self-esteem: Two dimensions of implicit self-esteem and their relation to narcissistic reactions. *Journal of Experimental Social Psychology*, *43*, 995–1001.

Schröder-Abé, M., Rudolph, A., Wiesner, A., & Schütz, A. (2007). Self-esteem discrepancies and defensive reaction to social feedback. *International Journal of Psychology*, *42*, 174–183.

Schwarz, N. (1990). Feelings as information: Informational functions of affective states. In E. T. Higgins, & R. M. Sorrentino (Eds.), *Handbook of motivation and cognition: Foundations of social behavior* (Vol 2, pp. 527–561). New York: Guilford Press.

Shimizu, M., & Pelham, B. W. (2004). The unconscious cost of good fortune: Implicit and explicit self-esteem, positive life events, and health. *Health Psychology*, *23*, 101–105.

Steiger, J. H. (1980a). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251.

Steiger, J. H. (1980b). Testing pattern hypotheses on correlation matrices: Alternative statistics and some empirical results. *Multivariate Behavioral Research*, *15*, 335–352.

Tracy, J. L., Cheng, J. T., Robins, R. W., & Trzesniewski, K. H. (in press). Authentic and hubristic pride: The affective core of self-esteem and narcissism. *Self and Identity.*

Vasu, E. S., & Elmore, P. B. (1975). The effects of multicollinearity and the violation of the assumption of normality on the testing of hypotheses in regression analysis. *Multiple Linear Regression Viewpoints*, *6*, 21–50.

Verplanken, B., Friborg, O., Wang, C. E., Trafimow, D., & Woolf, K. (2007). Mental habits: Metacognitive reflection on negative self-thinking. *Journal of Personality and Social Psychology*, *92*, 526–541.

Watson, D. (1988). Intraindividual and interindividual analyses of positive and negative affect: Their relation to health complaints, perceived stress, and daily activities. *Journal of Personality and Social Psychology*, *54*, 1020–1030.

DOI: 10.1002/per