Scandinavian Journal of Psychology, 2015, 56, 132–139 DOI: 10.1111/sjop.12189

Cognition and Neurosciences

A Swedish validation of the Berlin Numeracy Test

MARCUS LINDSKOG, NEDA KERIMI, ANDERS WINMAN and PETER JUSLIN

Uppsala University, Uppsala, Sweden

Lindskog, M., Kerimi, N., Winman A. & Juslin, P. (2015). A Swedish validation of the Berlin Numeracy Test. *Scandinavian Journal of Psychology, 56*, 132–139.

Recent research has highlighted the importance of considering an individual's level of *numeracy*, that is their numerical abilities, in a vast variety of judgment and decision making tasks. To accurately evaluate the influence of numeracy requires good and valid measures of the construct. In the present study we validate a Swedish version of the Berlin Numeracy Test (Cokely, Galesic, Schulz, Ghazal & Garcia-Retamero, 2012). The validation was carried out on both a student sample and a sample representative of the Swedish population. The Swedish BNT showed sound psychometrical properties in both samples. Further, in both samples the BNT had satisfactory convergent and discriminant validity when correlating with other measures of numeracy, while not being significantly related to measures of personality. With respect to predictive validity the results indicated divergent patterns in the two samples. In the student sample, participants scoring highest on the BNT outperformed those in the other three levels, which did not differ in performance. In contrast, in the population sample participants scoring lowest on the BNT performed worse than those in the other three levels, which did not differ in performance. Taken together, however, the results suggest that the Swedish version of the BNT should be considered a valid measure of numeracy in both Swedish student and population representative samples.

Key words: Berlin Numeracy Test, statistical numeracy, individual differences, decision making, student sample, population sample, Swedish validation.

Marcus Lindskog, Department of Psychology, Uppsala University, PO Box 1225, SE-751 42 Uppsala, Sweden. Tel. +46(0)18-471 21 04;
e-mail: marcus.lindskog@psyk.uu.se

INTRODUCTION

Modern society is becoming increasingly more numerate, requiring people to understand, evaluate, and act upon vast amounts of numerical information on a daily basis. *Numeracy*, the numerical equivalent to literacy, is an important factor in a wide range of decisions and judgments that include numbers (e.g., Lipkus & Peters, 2009; Reyna, Nelson, Han & Dieckmann, 2009). Consequently, researchers have tried to develop scales that can accurately capture the level of numeracy in decision makers (e.g. Cokely, Galesic, Schulz, Ghazal & Garcia-Retamero, 2012; Fagerlin, Zikmund-Fisher, Ubel, Jankovic, Derry & Smith, 2007; Lipkus, Samsa & Rimer, 2001; Weller, Dieckmann, Tusler, Mertz, Burns & Peters, 2013). A recent contribution is the Berlin Numeracy Test (BNT; Cokely *et al*., 2012), which is a short, adaptive test with good psychometrical properties. In this paper, we test the validity of the BNT in Swedish for two samples, one consisting of undergraduate students and one consisting of participants representative of the Swedish population.

What is numeracy?

The last twenty years have seen a growing interest in people's ability to understand, evaluate, and use numerical information (summarized in the concept of *numeracy*). Even though the literature contains several attempts to describe what numeracy is (for a review see Reyna *et al*., 2009) there is no real consensus on how it should be defined. It is, however, often described as the numerical version of literacy or quantitative literacy and captures the ability to process basic probability and numerical concepts (Lipkus *et al*., 2001). This ability is conceptualized as a continuous individual-difference variable that ranges from very low to very high (Lipkus & Peters, 2009).

Numeracy in judgment and decision making

One reason for the growing interest in numeracy is the documented relationship between numeracy and the ability to make good and informed decisions related to health issues (e.g., Nelson & Reyna, 2007; Peters, Hibbard, Slovic & Dieckmann, 2007) with lower numeracy related to poorer decisions. The modern patient often takes an active role in medical decisions, rather than relying on a decision from a health provider, and it is important that patients can understand the information, often framed in a numerical format (e.g., there is a 50% chance to survive, or 100 out of 1000 people recover), that is presented to them (Nelson, Reyna, Fagerlin, Lipkus & Peters, 2008). Accordingly, the link between the quality of judgments and decisions and an individual level of numeracy was initially studied within the area of health communication, including the sub-areas of health information, risk communication, and decision making in health settings (Reyna *et al*., 2009).

The insight that decisions might be influenced by an individual's level of numeracy has also resulted in a growing interest to study the concept in a more classical framework of judgment and decision making (e.g., Peters, Slovic, Västfjäll & Mertz, 2008). This research has shown that the relationship between numeracy, and judgment- and decision-quality is not limited only to the health domain. Rather, an individual's level of numeracy seems to be of importance for judgments and decisions in general. For example, people low on numeracy are more sensitive to framing effects (Peters, Västfjäll, Slovic, Mertz, Mazzocco & Dickert, 2006; Reyna *et al*., 2009), give less accurate estimates of risk (Black, Nease & Tosteson, 1995), and tend to ignore sample size information to a larger extent (Obrecht, Chapman & Gelman, 2009) than people with high numeracy. These findings suggest that numeracy is an important part of everyday judgment and

decision making and stress the importance of reliable scales with which to measure it.

Measuring numeracy

There have been several attempts to develop scales that quickly and efficiently measure an individual's level of numeracy (e.g. Cokely *et al*., 2012; Fagerlin *et al*., 2007; Lipkus *et al*., 2001; Weller *et al*., 2013). An initial scale introduced by Schwartz, Woloshin, Black, and Welch (1997), included only three items and was intended to screen patients' ability to evaluate the benefits of mammography. This scale was extended with eight additional items by Lipkus *et al*. (2001) to yield the, to date, most widely used scale (Expanded Numeracy scale, ENS). The ENS measures an individual's ability to convert probabilities into percentages (and vice versa), estimate probabilities, and convert frequencies into probabilities. The ENS was developed using a "highly educated sample" (Lipkus *et al*., 2001, p. 37) that showed a surprisingly low ability to correctly answer the items. For example, only 55% of participants could correctly answer the question "Imagine that we rolled a fair, six-sided die 1,000 times. Out of 1,000 rolls, how many times do you think the die would come up even (2, 4, or 6)?" Even though Lipkus *et al*. (2001) described their participants as highly educated, results from subsequent research indicate that the performance of participants in this study may have been atypically low. Several studies have shown considerably better results on the ENS for student participants (e.g., Peters *et al*., 2006; Peters & Levin, 2008) with performance coming close to ceiling. Even though performance on the ENS seems to depend on the population from which participants come, it has still been able to explain a considerable amount of individual differences on diverse judgment and decision making tasks (e.g., Peters *et al*., 2006; Schley & Peters, 2014).

Recently, a growing body of research has highlighted that the psychometric properties of the ENS are not optimal, especially when the participants come from a student population (e.g., Cokely *et al*., 2012; Peters *et al*., 2006; Peters & Levin, 2008; Weller *et al*., 2013). This has led to the development of new scales, more appropriate for use with student samples. For example, Peters and colleagues (2007) developed an extended version of the ENS where four additional items were included to increase the discriminability of the scale. Adding questions, however, also adds to the time it takes to complete the scale. Moreover, Cokely and colleagues (2012) showed that many of the existing numeracy scales lose their predictive power when variables such as intelligence are controlled for.

The Berlin Numeracy Test. In an attempt to develop a short and psychometrically sound test of statistical numeracy and risk literacy, Cokely *et al*. (2012) introduced the Berlin Numeracy Test (BNT). The test was intended to be short and to have increased discriminability as compared to the ENS.

The BNT can be carried out either as a traditional pen and paper test or as an adaptive test. The adaptive version presents the four questions in an adaptive structure. In the adaptive structure, which question will be answered next depends on whether the answer on the previous question was correct or not. The Appendix shows both the original English items, the translated Swedish items used for this study, and the adaptive structure of the BNT. The BNT has been translated into several different languages (including German, English, and Spanish) and has shown robust psychometrical properties. Furthermore, Cokely and colleagues (2012) showed that the BNT is correlated with the ENS and measures of cognitive ability, and not correlated with unrelated constructs such as agreeableness. Furthermore, they showed that the BNT scale has a better predictive power as compared to the ENS. Even though the BNT has proven to be useful even when presented in different languages (e.g., German), its validity in Swedish has not yet been established. In this paper, we will validate the BNT in Swedish.

The present study

Previous research suggests that numeracy is an important factor in many decisions. The BNT scale is the latest contribution to the list of numeracy scales. However, as with all measures of individual differences, it is important to ascertain that the BNT is valid for the population at hand. Especially because previous research has indicated that there might be cultural differences in level of numeracy (Cokely *et al*., 2012). The BNT was initially developed for individuals with an educational background corresponding to at least undergraduate university students. In Study 1, we therefore validate the Swedish BNT using a sample of university students. Even though a lot of research on numeracy is conducted using undergraduate students it is also a concept of interest for a wider range of participants, some of whom may not have a university-level education. It is reasonable to assume that an individual's level of numeracy might be influenced by their level of education, a possibility that might make the BNT unsuitable for other samples than those composed of students. In Study 2, we therefore use a representative sample of the Swedish population to investigate the validity of the BNT in a sample that is more heterogeneous than a student sample. Validating the scale on different samples adds to the generalizability of the scale.

STUDY 1: VALIDATION USING A STUDENT SAMPLE

The BNT was developed for educated populations. Therefore, in Study 1 we use university students to validate BNT. To validate the BNT, we investigate the distribution of performance, convergent validity, discriminant validity, and predictive validity. Moreover, we examined the relation between participants' BNT scores and their level of education.

Method

Participants. The student sample consisted of 123 participants. To be included in the study we required participants to have studied at least one semester at a university level. Two participants failing to meet this requirement were excluded from the data analysis. Of the remaining 121 participants, 46.3% were males. The ages of the participants were between 19 and 44 (M = 24.9, SD = 4.3). They were recruited from notice boards at university departments, online notice boards for research

participant recruitment, and online social networks. Participants were reimbursed by having a 6% chance to win five cinema tickets or a voucher worth 500 SEK in a lottery.

Materials. Materials in this study consisted of the BNT and two criterion-validity questions (Cokely *et al*., 2012), the Expanded Numeracy scale (Lipkus *et al*., 2001), the Subjective Numeracy scale (SNS; Fagerlin *et al*., 2007) and the Agreeableness subscale of the Big 5 personality scale (Bäckström, Björklund & Larsson, 2009). Each test is described in more detail below. In addition to the four measures, participants answered a battery of questions unrelated to the present study and thus not further reported here.

BNT. This BNT consists of four questions, and comes in two versions. One version is administered as a traditional paper and pen test, where participants answer all four questions, and one computer administered adaptive version, in which participants solve different questions depending on their past success in answering previous questions. In this study, we use the adaptive version. In the adaptive version participants answer 2–3 questions depending on their performance. The adaptive structure adjusts the difficulty of the subsequent questions based on the prior performance of the participant and is constructed to make all questions have about a 50% probability of being answered correctly. The test assigns participants to one of four skill-levels (1–4) of numeracy. The questions for the BNT (see Appendix) were translated into Swedish by the first and second author and re-translated into English for a check of consistency by a member of the Department of Psychology at Uppsala University who was naive to the original English version.

BNT criterion-validity questions. To measure the BNT's predictive validity, Cokely and colleagues (2012) developed criterion-validity questions. These questions, framed in the health care field, measure the understanding of everyday risk. In the present study we used two questions with a similar structure and contents to those used in Cokely *et al*. (2012) to investigate the predictive validity of the BNT. The choice of questions was motivated by the possibility to evaluate predictive validity similarly to previous studies. The questions, adapted from Cokely *et al*. (2012) are included in the Appendix in their English version. Participants, however, completed Swedish versions of the two questions. One of the questions is about medication of a drug, and the other is about mammography screening. Each question is followed by five statements where only one statement is correct and participants are asked which statement that is most useful when assessing the benefits of the medication/mammography.

Expanded Numeracy scale (ENS). The expanded numeracy scale (Lipkus *et al*., 2001) is the most widely used test of numeracy. Therefore, to investigate the validity of the BNT the predictive power of BNT will be compared to the this scale. The scale consists of 11 questions and is developed for highly educated populations. Three of the questions are taken from Schwartz *et al*. (1997), seven of the questions are framed in the health domain, and one question is for practice. All questions are open-ended except two, which have multiple-choice options.

Subjective Numeracy scale (SNS). The subjective numeracy scale (Fagerlin *et al*., 2007) consists of a set of eight questions developed to measure numeracy without a math test. Thus, the questions have participants rate their proficiency with numbers and calculations and their inclination to use numbers rather than words when describing numerical information in everyday events.

Agreeableness. Agreeableness encompasses the ability to be helpful, empathic, and trustworthy. Earlier studies on numeracy have shown that numeracy is unrelated to the personality trait Agreeableness (Cokely *et al*., 2012). Therefore, to investigate the discriminant validity of the BNT, we included ten items from the Big 5 questionnaire that measures Agreeableness.

Procedure. Participants received an email with a link to the online version of the tests and had one week to finish them. Before starting the tests, participants had to read and accept a consent form. Furthermore, participants were instructed to do the tests individually and not use calculator or other tools that could aid them in answering the questions. They were, however, allowed to use paper and pen. The order of the tests was counterbalanced. After conducting the tests, participants answered a number of demographic questions.

Results and discussion

In both Study 1 and Study 2, all results are calculated using the Swedish version of the BNT. For convenience, however, we use BNT as an abbreviation for both the English and Swedish versions of the test unless the context makes distinguishing between them difficult. Descriptive data for the three numeracy measures in both Study 1 and Study 2 are summarized in Table 1.

Quartile distribution. The BNT is designed to divide participants into four quartiles. We thus expected one quarter of participants in each of the four levels of numeracy (1–4). In line with this prediction, the results indicated an even distribution of participants over the four levels (L1: 20.7%, L2: 24.0%, L3: 21.5%, L4: 33.9%). There is a slight negative skew in the distribution with a higher proportion of participants in the fourth level than expected. However, the distribution did not differ significantly

Table 1. *Descriptive data for the three numeracy measures (Berlin Numeracy Test (BNT), Expanded Numeracy Scale (ENS), and Subjective Numeracy Scale (SNS)) in the student and population sample respectively*

	Sample					
	Student			Population		
Descriptive measure	BNT	ENS	SNS	BNT	ENS	SNS
Mean	2.69	9.94	4.08	2.47	9.41	3.94
Median	3.00	10.00	4.25	2.00	10.00	4.00
Standard deviation	1.15	1.60	0.87	1.03	2.04	0.97
Skew	–0.20	–2.5	–0.43	0.13	–2.0	–0.40
25^{th} perc.	2.0	10.0	3.5	2.0	9.0	3.3
75^{th} perc.	4.0	11.0	4.6	3.0	11.0	4.6

from a uniform distribution, $\chi^2(3, N = 121) = 5.38$, $p = 0.15$. Further, the adaptive structure of the BNT was designed to give an approximate median split after the first question (Cokely *et al*., 2012). That is, 50% of the participants should answer the first question correctly while 50% should not. The first question was answered correctly by 55.4% of participants, a proportion that did not differ significantly from the expected 50%, $t(121) = 1.40$, $p = .24$.

Convergent validity. As discussed above, several scales have been developed to measure numeracy. A measure claiming to tap this construct should thus show convergent validity and be related to other such measures. We therefore calculated the Pearson correlations between the BNT and the ENS and between the BNT and the SNS. These analyses showed that both the ENS, $r(119) = 0.32$, $p < 0.001$, and the SNS, $r(119) = 0.41$, $p < 0.001$, were significantly related to the BNT. This indicates that the BNT taps the same underlying construct as the ENS and the SNS.[1]

Discriminant validity. Numeracy is considered an individual ability unrelated to personality traits. Thus, to evaluate the discriminant validity of the BNT, we calculated the Pearson correlation between participants' level of numeracy as measured by BNT and their score on the Agreeableness subscale of the Big 5 personality scale. The analysis showed a non-significant correlation, $r(119) = –0.05$, $p = 0.59$, indicating the predicted discriminant validity.

Predictive validity. Numeracy is thought to be related to an individual's ability to solve problems that include numerical information, and several studies have shown that level of numeracy can predict performance in a range of such tasks (e.g., Peters *et al*., 2006). A numeracy scale should therefore be able to predict performance in tasks that include numerical information. Consequently, to test the ability of the BNT to predict such performance, we used two criterion-validity questions (Medicine and Mammography) related to medical risks. We combined the results from the two questions into a composite measure with the number of correct answers (0, 1, or 2) as a measure of performance on the criterion validity questions. The distribution of these scores was, 0: 10%, 1: 60%, and 2: 30% respectively and the mean scores in each of the four BNT-levels were; L1: 1.0, L2: 1.2, L3: 1.2, L4: 1.3. To investigate the predictive power of the BNT, we entered the score on the criterion-validity question composite as dependent variable and the BNT-level as independent variable into a one-way ANOVA. The analysis revealed that the effect of BNT on the criterion composite measure was not significant, $F(3, 117) = 1.3$, $p = 0.28$, which suggests that the BNT was not able to predict performance in the criterion-validity questions. The distribution of scores on the composite measure indicates that there might be a difference in difficulty between the two questions. Indeed, while 87% of participants answered the Mammography question correctly, only 33% answered the Medicine question correctly. Because of this, separate ANOVAs for the two questions were carried out. The proportion of participants in each of the four BNT-levels answering each of the two criterion validity questions correct is summarized in Table 2. These analyses indicated a significant effect of BNT on the Mammography question, $F(3, 117) = 2.8$, $p = 0.04$, but not on the Medicine question, $F(3, 117) = 1.3$, $p = 0.29$. Follow up analyses indicated that participants in L4 significantly outperformed participants in L1 and L2 on the Mammography question. Thus, while performance on the composite score could not be predicted by the BNT it was possible to predict performance on the Mammography question.

Table 2. *Proportion of participants giving correct answers on the criterion-validity questions (Medicine and Mammography), used to evaluate predictive validity, in each of the four BNT–levels (L1, L2, L3, L4) in the student and population sample, respectively*

	Criterion–validity question							
	Medicine				Mammography			
	BNT-level				BNT-level			
Sample	L1	L2	L3	L4	L1	L2	L3	L4
Student	0.20	0.45	0.35	0.32	0.80	0.76	0.88	0.98
Population	0.26	0.35	0.33	0.30	0.59	0.85	0.74	0.84

Unique predictive power. There have been several attempts to develop scales that measure numeracy. To motivate a new scale, like the BNT, it is thus not enough to have predictive power on its own. The scale should have unique predictive power over and above other numeracy scales. To investigate whether BNT has unique predictive power over the ENS, an ANCOVA with the composite criterion score as dependent variable, BNT as between-subjects independent variable and ENS as covariate was conducted. The analysis indicated that the effect of BNT on the composite criterion score was unaltered when entering ENS as covariate. Analyzing the Mammography question separately with a corresponding ANCOVA revealed that including ENS as covariate reduced the effect of BNT to non-significant $F(3, 116) = 2.2$, $p = 0.096$).[2]

Predictive power of the first question. The first question of the BNT is expected to give an approximate median split of participants. Cokely and colleagues (2012) thus argue that those who answer the first question of the BNT correctly belong to the top half of the educated participants. As described earlier, 55.4% of our participants answered the first question correct. To investigate if the first question could be used to predict performance on the criterion questions, we compared performance on the composite criterion score for those answering the first question correct with that of those answering it incorrect. This analysis indicated a non-significant difference between the two groups, $t(119) = 1.4$, $p = 0.15$. Analyzing the two questions separately indicated a significant difference for the Mammography question $t(119) = 2.7$, $p = 0.008$, but not for the Medicine question, $t < 1$. This indicates that the first question of the BNT has similar predictive properties as does the entire test.

BNT and education. As one of the demographic questions, participants reported the number of semesters they had studied at the university. To investigate if there is a relationship between the length of university education and numeracy, we calculated

the Pearson correlation between these two measures. The analysis revealed a non-significant correlation, $r(119) = 0.03$, $p = 0.76$. Participants also reported the subject that they majored in at the university. The participants major subject was coded by two independent coders (98% inter coder agreeability) into those that require the use of mathematics and/or statistics to a large extent (MS: e.g. physics, computer science, economics, etc.) and those that do so to a lesser extent (no-MS: e.g. psychology, physiotherapy, medicine, etc.). This coding resulted in 81 participants coded as no-MS and 37 as MS while three participants did not report their major subject. Comparing the level of numeracy of the MS and no-MS groups revealed a marginally significant difference, $t(116) = 1.8$, $p = 0.08$, with slightly better performance in the MS ($M = 3.0$, $SD = 1.2$) than in the no-MS group ($M = 2.6$, $SD = 1.1$).

STUDY 2: VALIDATION USING A POPULATION-BASED SAMPLE

The BNT was developed, and mainly validated, for university student populations (Cokely *et al*., 2012). University student populations are expected to be highly educated in the sense that they have taken at least one college course. In psychological studies most often these courses are in psychology, because of the availability of these students to researchers and due to course requirements. Even though a lot of research in judgment and decision making is carried out using participants from populations that are similar to such a student population, the general concept of numeracy is not confined within this population. The first numeracy scales (e.g. Schwartz *et al*., 1997) and indeed some developed later (e.g., Fagerlin *et al*., 2007), were intended to measure the level of numeracy of a much broader population than those found at universities. If it is possible to use the BNT as a valid measure of numeracy also in samples where participants are not as highly educated as university students, and where knowledge of statistics might be more heterogeneous, it would have two major benefits. First, the generalizability of the BNT would be much greater. Second, it would be possible to use the BNT also when conducting research outside of educational institutions. Because the BNT was developed using students it is an empirical question whether it will exhibit the properties required of a valid measure of numeracy when used on a less educated and more heterogeneous sample. Further, because there have been few attempts to use the BNT with groups other than students it is not obvious what to expect of the measure *a priori*. On the one hand it is possible that the skill set needed to correctly solve the more difficult items is acquired only through higher education and that a population-based sample would exhibit a strong skew in the distribution of the four levels. On the other hand it might be that a more general skill set acquired earlier in the school system is sufficient. If this is the case we would expect similar properties of the BNT in a population-based sample as in the student sample reported above. In Study 2, we investigate the validity of the BNT using a population-based sample. This was done in order to, if possible, extend the use of the BNT as a valid measure of numeracy from student samples to less educated samples.

Method

Participants. The Swedish population-based sample consisted of 227 participants (60.8% female) with ages between 21 and 62 ($M = 39.3$, $SD = 11.6$). Of these, three participants did not conduct the ENS and 23 participants did not answer the criterion validity questions of the BNT. Furthermore, 11 participants did not fully conduct the Big five. All participants were recruited by a random draw from a database where all Swedish residents are registered. Participants were reimbursed with the choice between a gift certificate valued 1,000 Swedish kronor (approximately $140) or the opportunity to give the same amount to charity.

Materials and apparatus. Materials relevant for this study was similar to the material used in Study 1 with the exception that we collected data on all of the Big 5 traits.

Design and procedure. Participants first completed the BNT and after approximately one hour, the ENS on a computer at the university facilities. For the BNT, participants had the option to use paper and pen. Between the BNT and the ENS, participants also participated in other tests, which are not relevant for this paper. One week after their participation, participants received a link to online versions of the Big 5 questionnaire, the two criterion validity questions for the BNT, and the SNS.

Results and discussion

Quartile distribution. In line with the predictions of the BNT, and the results of Study 1, there was an even distribution of participants over the BNT levels (L1: 19.4%, L2: 35.2%, L3: 24.2%, L4: 21.1%). However, chi-square analysis showed that the distribution deviated significantly from a uniform distribution, $\chi^2(3, N = 227) = 13.79$, $p = 0.003$, indicating a slight positive skew in the distribution. The first question was answered correctly by 46.1% of participants, a proportion that did not deviate significantly from the expected 50%, $t(220) = 1.1$, $p = 0.25$. Thus, despite the slight positive skew the first question was still able to give the intended approximate median split.

Convergent validity. Similar to the student based sample in Study 1, the Pearson correlations showed a significant positive correlation between the BNT and the ENS, $r(222) = 0.47$, $p < 0.001$, and between the BNT and the SNS, $r(214) = 0.35$, $p < 0.001$, indicating convergent validity.[3]

Discriminant validity. In the population-based sample, Pearson correlations showed non-significant correlations between the BNT and Agreeableness, $r(214) = 0.02$, $p = 0.83$, Conscientiousness, $r(214) = –0.10$, $p = 0.15$, Stability, $r(214) = 0.07$, $p = 0.33$, Openness, $r(214) = 0.09$, $p = 0.20$, and Extraversion, $r(214) = –0.04$, $p = 0.55$, suggesting discriminant validity.

Predictive validity. Similar to Study 1, there was a difference in difficulty between the two questions used to test criterion validity with 31% of participants answering the Medicine question correct while 77% of participants answered the Mammography question correct. Thus, as in Study 1, we analyzed the predictive

power of the BNT using a composite measure of both criterion questions, and for each question separately. When the criterion questions were combined, the distribution of the scores were: 0: 13.6%, 1: 53.5%, and 2: 21.1% respectively and the mean scores in each of the four BNT-levels were; L1: .85, L2: 1.2, L3: 1.1, L4: 1.1. As in Study 1 we investigated the predictive power by means of a one-way ANOVA. This analysis, revealed a significant effect of BNT on performance in the composite criterion score, $F(3, 197) = 2.85$, $p = 0.04$. LSD post hoc analysis revealed that participants in L1 performed significantly worse than those in L2 ($p = 0.005$) and L4 ($p = 0.03$) but not those in L3 ($p = 0.1$). All other comparisons had $p > 0.27$. Analyzing the Medicine and Mammography questions separately showed a significant effect of BNT on performance for the Mammography question, $F(3, 197) = 3.8$, $p = 0.01$, but not for the Medicine question, $F < 1$ (see Table 2 for the proportion of participants in each of the four BNT-levels answering each of the two criterion validity questions correct). Thus, while the results in Study 1 indicated that the BNT could be used to predict which participants that would successfully answer the Mammography question (those in L4) the results in Study 2 showed that the BNT could be used to predict which participants that would fail to answer the criterion questions correctly (those in L1). These results indicate that while the BNT has some predictive validity, its predictive properties will differ over populations.

Unique predictive power. The unique predictive power of the BNT over the ENS was investigated by means of an equivalent ANCOVA to that of Study 1. The analysis indicated that including ENS as covariate eliminated the effect of BNT on the composite criterion score, $F(3, 196) = 1.76$, $p = 0.16$. Thus, the BNT did not have predictive power over and above the ENS. For the Mammography question the corresponding ANCOVA resulted in the effect of BNT becoming marginally significant, $F(3, 196) = 2.4$, $p = 0.07$.[4]

Predictive power of the first question. We investigated the possibility that the first question of the BNT could predict performance on the composite criterion score by comparing performance for those answering the first question correct with those answering it incorrect. The analysis showed that there was no difference between the two groups, $t(196) = 0.4$, $p = 0.7$, suggesting that the first question could not be used to predict performance on the composite criterion score.[5]

BNT and education. The population sample was more heterogeneous with respect to education than the student sample. This gives the possibility to investigate a possible relationship between BNT and education in greater detail than in the student sample. Participants (4.5%) that did not indicate their level of education were excluded from the analysis. The remaining participants were divided into four groups based on their level of education. The first group included participants with only middle-school education (2.2%), while the second group consisted of those with only occupational training or high-school education (30.4%). The third included participants with undergraduate university level studies (38.3%), and the fourth group consisted of those with graduate studies (24.6%). A Pearson correlation indicated a significant positive correlation between level of education and numeracy $r(217) = 0.22$, $p < 0.001$ with the more educated also being more numerate.

GENERAL DISCUSSION

Previous research has indicated that numeracy is an important individual difference factor in severe judgment and decision making tasks (e.g., Peters *et al*., 2006). Even though several scales that measure numeracy have been developed, none of them have, to our knowledge, been validated in Swedish. The aim of the present study was, therefore, to perform such a validation for one of the currently available measures, the Berlin Numeracy Test (Cokely *et al*., 2012). The validation was conducted on two separate samples. In Study 1 we used a sample of university students. This is the type of sample that the BNT was originally developed for. Validly measuring numeracy is, however, not only of interest in student populations. To further extend the possible use of the Swedish version of the BNT we also conducted a validation using a sample of participants that were representative of the Swedish population. We validated the BNT by investigating the distribution of responses in addition to the convergent, discriminant, and predictive validity of the test.

The first question of the BNT is intended to give an approximate median split of the participants. This was found in both the population and student sample. In addition the distribution over the four levels of performance should be uniform (i.e., approximately 25% of participants should be found on each level). While the uniform distribution was found in the student sample there was a slight positive skew in the population sample. This was expected because the general level of numeracy should be higher in the student population than in the general population. Thus, when using the BNT with samples that are more heterogeneous than student samples some caution is advised due to the slight skew. It should be noted, however, that even though the distribution deviates significantly from the uniform it is still very well behaved compared to other measures of numeracy (e.g. Lipkus *et al*., 2001) that exhibit strong negative skew even in population samples.

In both samples we found that the BNT exhibited the predicted convergent and discriminant validity when correlating significantly with the ENS (Lipkus *et al*., 2001) and the SNS (Fagerlin *et al*., 2007) while not being significantly related to measures of personality.

Because numeracy is considered an important factor that influences performance in judgment and decision making tasks it was expected that it would be possible to predict performance on the two criterion-validity questions from performance on the BNT. Further, it was expected that this would be possible over and above the predictive power of the ENS. In both studies one of the two criterion-validity questions (Medicine) turned out to be difficult with only 30% of participants answering it correctly. The BNT was not able to predict performance on this question in either of the two samples. It was, however, possible to predict performance on the other (Mammography) of the two questions. The results indicated divergent patterns of predictability in the two samples. In the student sample, participants scoring highest on the BNT (i.e., L4) outperformed those in the other three levels, which did not differ in performance. In contrast, in the population sample participants scoring lowest on the BNT (i.e. L1) performed worse

than those in the other three levels, which did not differ in performance. This might suggest that the predictive ability of the BNT might differ in different populations. Taken together, these results deviate from results found with other versions of the BNT (Cokely *et al*., 2012) and might indicate that the Swedish version of the BNT does not have the required predictive abilities. One of the real values of numeracy is in predicting decision behaviors. A limitation of the present study is that we only used a limited set of criterion-validity questions. While it might had been beneficial to include a large number of tasks to evaluate predictive validity it is not necessarily easy to choose which ones that should have been included. Even though numeracy should conceptually be able to predict judgments and decisions it is not the case that it is an equally strong predictor for all judgments and decisions (see e.g., Winman, Juslin, Lindskog, Nilsson & Kerimi, 2014). Therefore, one reason for the issues with low predictive power in the present study could be that we only used two criterion-validity questions and that the difficulty of these two questions were not optimal to find predictive validity. Another possibility is that numeracy is not a good predictor for the specific questions used here. The choice of questions in the present study was, however, motivated by the types of questions used to validate the BNT in other samples (Cokely *et al*., 2012). It should be noted, however, that previous studies (e.g., Winman *et al*., 2014) using the same Swedish translation of the BNT have indicated that level of numeracy as measured with the BNT is an important individual difference factor in other judgment and decision making tasks (e.g. overconfidence, linearity of calibration curves, and rate of conjunction fallacies). This indicates that even though the predictive validity of Swedish version of the BNT was not fully as expected in the present study, it does predict performance on a set of diverse judgment and decision making tasks. It will be an important issue for future research to address the issue with predictive validity of the Swedish BNT in further detail.

In both studies we investigated the relationship between education and BNT. The results from the population sample in Study 2 indicated that more education was associated with better performance on the BNT. However, the results from Study 1 indicated that the number of completed semesters at university did not correlate with performance on the BNT. Only after we had separated the participants in the student sample into those taking courses requiring a lot of mathematics and/or statistics and those taking courses with no such requirements could we find a marginally significant difference. Taken together, these results indicate that if level of education has an influence on numeracy it is primarily education prior to undergraduate university education that is of importance.

To summarize; the Swedish version of the BNT exhibited good psychometric properties in both a student and a population representative sample. While the test showed satisfactory convergent and discriminant validity there was some concerns with predictive validity. Taken together, however, the Swedish version of the BNT should be considered a valid measure of numeracy in both Swedish student and population representative samples.

This research was sponsored by the Swedish Research Council. The authors are indebted to Anja Löfgren and Johan Eklund for help with the data collection.

NOTES

[1] Weller *et al*. (2013) suggested a numeracy scale consisting of five items from the ENS and three additional items. Although our participants did not complete the three additional items we calculated the correlation between the five ENS items from Weller *et al*. (2013) and the BNT. The analysis indicated, similar to the correlations with ENS and SNS, that BNT was significantly related ($r(119) = 0.29$, $p = 0.001$) to the five-item scale.

[2] We also tested the predictive power of ENS by itself on the two criterion-value questions by means of three separate simple-regression analyses. The results indicated that while ENS did not significantly predict performance on the Medicine question ($\beta = 0.05$, $p = 0.53$) it could predict performance both on the Mammography question ($\beta = 0.16$, $p = 0.03$) and the composite measure ($\beta = 0.14$, $p = 0.046$).

[3] The BNT was also significantly related ($r(218) = 0.48$, $p < 0.001$) to the five item scale suggested by Weller *et al*. (2013).

[4] We also tested the predictive power of ENS by itself on the two criterion-value questions by means of three separate simple-regression analyses. The results indicated that ENS could not significantly predict performance on the Medicine question ($\beta = 0.05$, $p = 0.61$), the Mammography question ($\beta = 0.14$, $p = 0.13$) or on composite measure ($\beta = 0.12$, $p = 0.20$).

[5] Analyzing the two questions separately showed the same pattern of results.

REFERENCES

Bäckström, M., Björklund, F. & Larsson, M. R. (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality*, *43*, 335–344.

Black, W. C., Nease, R. F. & Tosteson, A. N. (1995). Perceptions of breast cancer risk and screening effectiveness in women younger than 50 years of age. *Journal of the National Cancer Institute*, *87*, 720–731.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S. & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision Making*, *7*, 25–47.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A. & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the subjective numeracy scale. *Medical Decision Making*, *27*, 672–680.

Lipkus, I. M. & Peters, E. (2009). Understanding the role of numeracy in health: Proposed theoretical framework and practical insights. *Health Education & Behavior*, *36*, 1065–1081.

Lipkus, I. M., Samsa, G. & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, *21*, 37–44.

Nelson, W. & Reyna, V. (2007). Numeracy: A critical (and often overlooked) competence for health decision making. *Annals of Behavioral Medicine*, *33*, S8–S8.

Nelson, W., Reyna, V. F., Fagerlin, A., Lipkus, I. & Peters, E. (2008). Clinical implications of numeracy: Theory and practice. *Annals of Behavioral Medicine*, *35*, 261–274.

Obrecht, N. A., Chapman, G. B. & Gelman, R. (2009). An encounter frequency account of how experience affects likelihood estimation. *Memory & Cognition*, *37*, 632–643.

Peters, E., Hibbard, J. H., Slovic, P. & Dieckmann, N. (2007). Numeracy skill and the communication, comprehension, and use of risk-benefit information. *Health Affairs*, *26*, 741–748.

Peters, E. & Levin, I. P. (2008). Dissecting the risky-choice framing effect: Numeracy as an individual-difference factor in weighting risky and riskless options. *Judgment and Decision Making*, *3*, 435–448.

Peters, E., Slovic, P., Västfjäll, D. & Mertz, C. K. (2008). Intuitive numbers guide decisions. *Judgment and Decision Making*, *3*, 619–635.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K. & Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, *17*, 407–413.

Reyna, V. F., Nelson, W. L., Han, P. K. & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin*, *135*, 943–973.

Schley, D. & Peters, E. M. (2014). Assessing "economic value": Symbolic-number mappings predict risky and riskless valuations. *Psychological Science*, *25*, 753–761.

Schwartz, L. M., Woloshin, S., Black, W. C. & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine*, *127*, 966–972.

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J. & Peters, E. (2013). Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. *Journal of Behavioral Decision Making*, *26*, 198–212.

Winman, A., Juslin, P., Lindskog, M., Nilsson, H. & Kerimi, N. (2014). The role of ANS-acuity and Numeracy for the calibration and the coherence of subjective probability judgments. *Frontiers in Psychology*, *5*, 851. doi:10.3389/fpsyg.2014.00851.

Received 30 June 2014, accepted 31 October 2014

APPENDIX

English items of the BNT (Cokely et al., 2012):

Q1. Out of 1,000 people in a small town 500 are members of a choir. Out of these 500 members in the choir 100 are men. Out of the 500 inhabitants that are not in the choir 300 are men. What is the probability that a randomly drawn man is a member of the choir? Please indicate the probability in percent. _____%

Q2a. Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws how many times would this five–sided die show an odd number (1, 3 or 5)? _____ out of 50 throws.

Q2b. Imagine we are throwing a loaded die (6 sides). The probability that the die shows a 6 is twice as high as the probability of each of the other numbers. On average, out of these 70 throws how many times would the die show the number 6? _____out of 70 throws.

Q3. In a forest 20% of mushrooms are red, 50% brown and 30% white. A red mushroom is poisonous with a probability of 20%. A mushroom that is not red is poisonous with a probability of 5%. What is the probability that a poisonous mushroom in the forest is red? _____

Swedish translation of items for the BNT:

Q1. Av 1 000 människor i en liten stad är 500 medlemmar i en kör. Av dessa 500 medlemmar i en kör är 100 män. Av de 500 invånarna som inte är med i en kör är 300 män. Vad är sannolikheten att en slumpmässigt dragen man är medlem i kören? Ange sannolikheten i procent. _____%

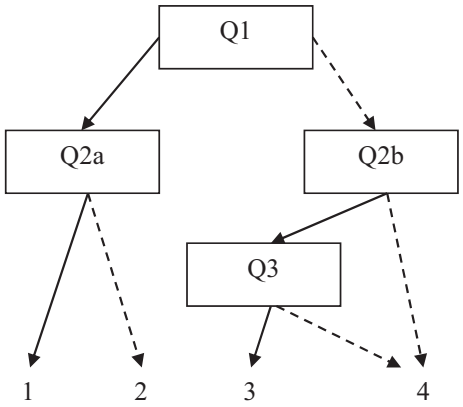
Q2a. Tänk dig att vi kastar en femsidig tärning 50 gånger. Av dessa 50 kast hur många gånger kommer denna femsidiga tärning till slut att visa en udda siffra (1, 3 eller 5)? _____av 50 kast.

Q2b. Tänk dig att vi kastar en falsk tärning (6 sidor). Sannolikheten att tärningen visar 6 är dubbelt så stor som sannolikheten för var och en av de andra siffrorna. Tänk dig nu att vi kastar tärningen 70 gånger. Av dessa 70 kast hur många gånger kommer tärningen till slut att visa siffran 6? _____ av 70 kast.

Q3. I en skog är 20% av svamparna röda, 50% är bruna och 30% är vita. En röd svamp är giftig med en sannolikhet av 20%. En svamp som inte är röd är giftig med en sannolikhet av 5%. Vad är sannolikheten att en giftig svamp i skogen är röd? Ange sannolikheten i procent. _____%

Adaptive structure of the BNT (Cokely et al., 2012).

Dashed arrows indicate the next question after a correct answer and solid arrows indicate the next question after an incorrect answer:



Questions used to test the predictive validity of the BNT:

Medicine. Gritagrel – a 50% reduction of strokes. Gritagel is a new medication to avoid strokes. People that took Gritagel showed only half the risk of stroke compared to people that took a placebo. Which information would be most helpful when estimate the usefulness of Gritagel:

A: How many people were in the Placebo group
B: How old were the participants of the study
C: The risk of stroke for people who took another medication for the same reason
D: Whether Gritagel has been recommended by professional doctors
E: The risk of having a stroke for people who don't take Gritagel

Mammography. Not every positive result of that test actually means that a woman does have breast cancer. Which of the following would be most helpful when estimating the benefits of mammography?

A: How much one screening costs.
B: How many women go to the screening.
C: How many women are treated for breast cancer.
D: How many women who get a positive result have breast cancer.
E: Has it been recommended by many doctors.