Behavioural and Cognitive Psychotherapy (2019), **47**, 559–572
doi:10.1017/S1352465819000018

BABCP
www.babcp.com

MAIN

The Negative Effects Questionnaire: psychometric properties of an instrument for assessing negative effects in psychological treatments

Alexander Rozental[1,2,*], Anders Kottorp[3], David Forsström[4], Kristoffer Månsson[1,5,6], Johanna Boettcher[7], Gerhard Andersson[1,8], Tomas Furmark[6] and Per Carlbring[5,9]

[1]Centre for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden, [2]Institute of Child Health, University College London, London, UK, [3]Faculty of Health and Society, Malmö University, Malmö, Sweden, [4]Department of Public Health Sciences, Stockholm University, Stockholm, Sweden, [5]Department of Psychology, Stockholm University, Stockholm, Sweden, [6]Department of Psychology, Uppsala University, Uppsala, Sweden, [7]Department of Clinical Psychology and Psychotherapy, Freie Universität Berlin, Berlin, Germany, [8]Department of Behavioural Sciences and Learning, Linköping University, Linköping, Sweden and [9]Department of Psychology, University of Southern Denmark, Odense, Denmark
*Corresponding author. Email: alexander.rozental@ki.se

(Received 16 November 2017; revised 19 September 2018; accepted 26 October 2018; first published online 15 March 2019)

Abstract

Background: Psychological treatments provide many benefits for patients with psychiatric disorders, but research also suggests that negative effects might occur from the interventions involved. The Negative Effects Questionnaire (NEQ) has previously been developed as a way of determining the occurrence and characteristics of such incidents, consisting of 32 items and six factors. However, the NEQ has yet to be examined using modern test theory, which could help to improve the understanding of how well the instrument works psychometrically.

Aims: The current study investigated the reliability and validity of the NEQ from both a person and item perspective, establishing goodness-of-fit, item bias, and scale precision.

Method: The NEQ was distributed to 564 patients in five clinical trials at post-treatment. Data were analysed using Rasch analysis, i.e. a modern test theory application.

Results: (1) the NEQ exhibits fairness in testing across sociodemographics, (2) shows comparable validity for a final and condensed scale of 20 instead of 32 items, (3) uses a rating scale that advances monotonically in steps of 0 to 4, and (4) is suitable for monitoring negative effects on an item-level.

Conclusions: The NEQ is proposed as a useful instrument for investigating negative effects in psychological treatments, and its newer shorter format could facilitate its use in clinical and research settings. However, further research is needed to explore the relationship between negative effects and treatment outcome, as well as to test it in more diverse patient populations.

Keywords: negative effects; Negative Effects Questionnaire; psychological treatments; Rasch analysis

Introduction

Psychological treatments can provide patients with effective means of overcoming mental distress and increasing their well-being (McHugh and Barlow, 2010). Research on the efficacy of evidence-based approaches, such as cognitive behaviour therapy (CBT), suggest that a large number of patients improve thanks to the interventions they receive (Hofmann et al., 2012). However, not everyone seems to benefit, with only half of the patients being regarded as responders at post-treatment and follow-up (Loerinc et al., 2015). Similarly, several investigations suggest that

a small proportion of all patients deteriorate during treatment (Boswell et al., 2015). For instance, Hansen et al. (2002) found that 8.2% fared worse in routine outpatient care, which can be compared with 6.9% within a psychiatric population (Mechler and Holmqvist, 2016), and 5.8% in clinical trials of internet-based CBT (Rozental et al., 2017).

Non-response and deterioration are, however, largely determined using different statistical procedures, cut-offs, or diagnostic criteria. Yet other negative effects might occur, but have thus far gained less attention (Castonguay et al., 2010). A major problem in relation to exploring such cases is the fact that no reliable instrument has existed, making it difficult to investigate their incidence and characteristics. Strupp and Hadley (1977) suggested that stigma, dependency and novel symptoms could occur and affect the patient negatively, but the means for their assessment were not discussed. A comprehensive rating system for videotaped sessions was later proposed, the Vanderbilt Negative Indicators Scale (VNIS; Suh et al., 1986), but it never gained widespread attention. Recently, attempts have instead been made to investigate negative events from the perspective of the patient, most notably the Inventory for the assessment of Negative Effects of Psychotherapy (INEP; Ladwig et al., 2014), and the Experiences of Therapy Questionnaire (ETQ; Parker et al., 2013). However, these instruments have not yet been used to a great extent for patients undergoing treatment. Furthermore, some of their items seem hard to apply in some settings, e.g. insurance issues do not target negative effects explicitly, or are more related to malpractice than unintended effects of evidence-based approaches.

Rozental et al. (2016) thus developed a novel instrument to assess negative effects in psychological treatments to overcome some of the previous shortcomings. Using the results from a study on negative effects in a clinical trial on social anxiety disorder (Boettcher et al., 2014), a consensus statement among researchers (Rozental et al., 2014), a qualitative analysis of the responses to open-ended questions among patients in treatment (Rozental et al., 2015), and a literature review, items for the Negative Effects Questionnaire (NEQ) were generated and investigated using an exploratory factor analysis. The resulting instrument consists of 32 items and six factors, explaining a total variance of 57.6%: symptoms, quality, dependency, stigma, hopelessness, and failure. Findings also indicate that symptoms accounted for 36.6%, possibly being the most important factor in terms of negative effects in psychological treatments, such as 'I had more problems with my sleep' (Item 1). Furthermore, one-third of the participants reported having experienced unpleasant memories, stress and anxiety, suggesting that these incidents could be fairly common in treatment.

However, an exploratory factor analysis assumes that items are scored on an interval or quasi-interval scale, in line with classical test theory (Wright, 1977). This presumes that all items are equally difficult for the respondent, or person, to complete, which might not be the case in reality. In addition, it does not allow a separation between persons and items, that is, to assess not only how well each item fits the underlying construct, but also the person's response patterns. This can be helpful for identifying abnormal responses that might warrant further development of the instrument (Andrich, 1978). Rasch analysis, on the other hand, which is based on modern test theory, can be applied in order to analyse ordinal data in a way that provides linear measures, thereby addressing some of the caveats of classical test theory (Wright, 1996). This generates estimates of reliability and validity both for persons and items, making it feasible to study the instrument with greater depth (Waugh and Chapman, 2004). Furthermore, Rasch analysis can specifically be applied to test the dimensionality of an instrument, which, in the case of negative effects, can be assumed to be unidimensional, i.e. forms a single underlying construct. This seems plausible from a theoretical point of view, that is, negative effects should constitute one type of outcome related to psychological treatments, yet has never been tested previously.

The purpose of the current study was therefore to use Rasch analysis to further examine the NEQ, based on data from patients having completed the instrument at post-treatment in five clinical trials ($N = 564$). This has not been applied before in relation to negative effects, something that could shed some light on the psychometric properties of an instrument that might become

useful in clinical and research settings. In particular, such a method makes it possible to detect item bias and to explore whether each item performs in a comparable way across sociodemographics. A similar study was conducted using Rasch analysis for the Depression, Anxiety and Stress Scales (Lovibond and Lovibond, 1995), suggesting that a number of items could be removed and that it was not supported as a general instrument for mental distress (Shea et al., 2009).

The aim of the current study was thus twofold: to explore the response categories of the NEQ to see if they are of incremental scale steps, i.e. 0–4, and to examine the response pattern and goodness-of-fit between persons and items. The overall objective was to determine the usefulness of the NEQ as a way of exploring negative effects in psychological treatments.

Method

Participants

Participants were recruited from five clinical trials of spider phobia, perfectionism, social anxiety disorder, and loneliness ($N = 564$). Each case involved self-referrals and the studies were advertised in Sweden via national and regional newspapers and radio shows, social media, posters and flyers. A complete overview of the sociodemographics and clinical variables at pre-treatment is given in Table 1. Because not every clinical trial requested the same type of information from the participants, there was some degree of systematic missing data, e.g. living with someone, prior psychological treatment, and prior or ongoing psychotropic medication. Also, due to publication issues, symptom severity was not possible to present for one of the clinical trials. In addition, one of the clinical trials was included as part of the exploratory factor analysis of the NEQ, i.e. social anxiety disorder ($n = 189$) (Rozental et al., 2016).

Treatment and therapists

The psychological treatments that were administered in the clinical trials consisted of CBT, delivered in various formats: face-to-face, virtual reality, and via the internet, with or without guidance from a therapist, or by support on demand (Andersson et al., 2017). The therapists were masters degree students having undergone basic clinical training or more experienced therapists in advanced clinical training (i.e. psychotherapists in training). As for the internet conditions, participants received weekly modules consisting of both reading material and exercises to be completed by the participants every week, comparable to a self-help book (Andersson, 2016). The psychological treatments ranged from one session to 9 weeks; shortest for spider phobia and longest for social anxiety disorder.

Procedure

The participants filled out their sociodemographics and several outcome measures during the recruitment process before being assessed for eligibility. This was performed on a secure online interface using an auto generated identification code, such as 1234abcd, thereby ensuring anonymity and minimizing data loss (Vlaescu et al., 2016). Upon completing their treatment, the participants answered the outcome measures again, with the addition of the NEQ (Rozental et al., 2016). The only exception was the clinical trial of spider phobia where paper and pencil was used.

Measures

The Negative Effects Questionnaire. The NEQ was developed by Rozental et al. (2016) with the aim of investigating the occurrence and characteristics of negative effects in psychological treatments. The process of developing the instrument is described in detail in the original study. The

Table 1. Sociodemographic characteristics, symptom severity at pre-treatment, type of treatment, and number of participants reporting negative effects in each clinical trial

	Spiderphobia (n = 100)	Perfectionism (n = 156)	Social anxiety disorder (n = 189)	Loneliness (n = 73)	Social anxiety disorder (n = 46)	
	1 session	8 weeks, 8 modules	6 weeks, 9 modules	8 weeks, 8 modules	9 weeks, 9 modules	Total (n = 564)
Gender: *n* (%)						
Male	83 (83.0)	20 (12.8)	43 (22.8)	21 (28.8)	17 (37.0)	117 (20.7)
Female	16 (16.0)	135 (86.5)	146 (77.2)	52 (71.2)	29 (63.0)	445 (78.9)
Non-binary	1 (1.0)	1 (0.6)	0 (0.0)	0 (0)	0 (0)	2 (0.4)
Age (years): mean (*SD*)	34.1 (10.4)	34.1 (9.1)	35.3 (12.5)	47.2 (17.6)	30.7 (8.3)	35.9 (12.6)
Civil status: *n* (%)						
Single	26 (26.0)	45 (28.8)	64 (33.9)	50 (68.5)	13 (28.3)	198 (35.1)[b]
Relationship	74 (74.0)	111 (71.2)	125 (66.1)	23 (31.5)	30 (65.2)[a]	363 (64.4)[b]
Children: *n* (% yes)	39 (39.0)	63 (40.4)	95 (50.3)	41 (56.2)	19 (41.3)[a]	257 (45.6)[b]
Living with someone: *n* (% yes)	76 (76.0)	–	134 (70.9)	25 (34.3)	30 (65.2)[a]	265 (47.0)[c]
Highest educational level: *n* (%)						
Elementary school	3 (3.0)	1 (0.6)	10 (5.3)	2 (2.7)	3 (6.5)	19 (3.4)
High school/college	33 (33.0)	35 (22.4)	73 (38.6)	17 (23.3)	9 (19.6)	167 (29.6)
University	63 (63.0)	115 (73.7)	104 (55.0)	53 (72.6)	34 (73.9)	369 (65.4)
Postgraduate	1 (1.0)	5 (3.2)	2 (1.1)	1 (1.4)	0 (0.0)	9 (1.6)
Employment: *n* (%)						
Unemployed	2 (2.0)	7 (4.5)	14 (7.4)	1 (1.4)	2 (4.3)	26 (4.6)
Student	16 (16.0)	37 (23.7)	45 (23.8)	15 (20.5)	19 (41.3)	132 (23.4)
Employed/self-employed	79 (79.0)	101 (64.7)	119 (63.0)	33 (45.2)	25 (54.3)	357 (63.3)
Retired	1 (1.0)	0 (0.0)	4 (2.1)	20 (27.4)	0 (0.0)	25 (4.4)
Parental leave	1 (1.0)	4 (2.6)	4 (2.1)	0 (0.0)	0 (0.0)	9 (1.6)
Sick leave	1 (1.0)	3 (1.9)	3 (1.6)	2 (2.7)	0 (0.0)	9 (1.6)
Other	0 (0.0)	4 (2.6)	0 (0.0)	2 (2.7)	0 (0.0)	6 (1.1)
Clinical severity mean *(SD)*						
Patient Health Questionnaire – 9 items	2.6 (3.5)	9.6 (5.5)	8.7 (4.8)	9.8 (5.0)	–	7.9 (5.5)[d]

(*Continued*)

Table 1. (*Continued*)

	Spiderphobia ($n = 100$)	Perfectionism ($n = 156$)	Social anxiety disorder ($n = 189$)	Loneliness ($n = 73$)	Social anxiety disorder ($n = 46$)	
	1 session	8 weeks, 8 modules	6 weeks, 9 modules	8 weeks, 8 modules	9 weeks, 9 modules	Total ($n = 564$)
Generalized Anxiety Disorder – 7 items	2.6 (3.1)	8.3 (5.1)	8.6 (4.5)	6.9 (4.5)	–	7.1 (5.0)[d]
Brunnsviken Brief Quality of Life Scale	76.0 (14.8)	42.7 (16.4)	32.2 (5.2)	32.4 (17.4)	–	43.8 (21.1)[d]
Prior psychological treatment *n* (% yes)	23 (23.0)	4 (2.6)	79 (41.8)	34 (46.6)	9 (19.6)	149 (26.4)
Prior or ongoing psychotropic medication *n* (% yes)	10 (10.0)	19 (12.2)	54 (28.6)	28 (38.4)	9 (19.6)	120 (21.3)
Psychological treatment *n* (%)						
Face-to-face	50 (50.0)	–	–	–	–	50 (8.9)
Virtual reality	50 (50.0)	–	–	–	–	50 (8.9)
Internet (guided)	–	78 (50.0)	–	36 (49.3)	46 (100.0)	160 (28.4)
Internet (unguided)	–	–	189 (100.0)	–	–	189 (33.5)
Internet (support on demand)	–	78 (50.0)	–	37 (50.7)	–	115 (20.4)
Reporting any type of negative effect caused by treatment*n* (% yes)	57 (57.0)	85 (54.5)	105 (55.6)	4 (5.5)	30 (65.2)	281 (49.8)

[a]Category not applicable in $n = 3$; [b]based on $n = 561$; [c]based on $n = 405$; [d]based on $n = 518$.

exploratory factor analysis resulted in a rotated factor-solution with 32 items and the following six factors: symptoms, quality, dependency, stigma, hopelessness, and failure. The NEQ was found to have a good internal consistency, α for the full instrument .95, range .72 to .93 for the six separate factors. The instrument also consists of one open-ended question in order to capture other negative effects that are not included among the items, but this was not explored in the current study.

Outcome measures

Each clinical trial included in the current study distributed a primary outcome measure selected by relevance, for instance the Spider Phobia Questionnaire (Muris and Merckelbach, 1996). Several secondary outcome measures were also administered; the Patient Health Questionnaire – 9 items (PHQ-9; Löwe et al., 2004), the Generalized Anxiety Disorder – 7 items (GAD-7; Spitzer et al., 2006), and the Brunnsviken Brief Quality of Life Scale (BBQ; Lindner et al., 2016). These are, however, only presented descriptively in Table 1 for an overview of the sample.

Statistical analysis

In order to investigate and evaluate the validity of the internal structure and response processes of the NEQ, Rasch analysis was applied, following the same steps as described in Lerdal et al. (2016). The software WINSTEPS, version 3.91.0.0, was used for all analyses, implementing a rating scale model as all of the items in the NEQ are scored on a similar rating scale category. Rasch analysis converts the patterns of raw scores from the NEQ into item and person equal-interval measures simultaneously, using a logarithmic transformation of the odds probabilities of the responses (Bond and Fox, 2013). This converted item measure is then applied to determine whether they are scored on a similar unidimensional construct, which is often viewed as crucial in terms of validity in both classical and modern test theory (Spector, 1992). In a similar manner, the converted person measure is utilized to evaluate person response validity and the precision of the scale.

The psychometric properties of the NEQ rating scale categories were initially examined using the following criteria: (a) minimum of 10 responses per step category, (b) the average measures for each step category should advance monotonically, and (c) outfit Mean Square (*MnSq*) values less than 2.0 for the step category calibrations (Linacre, 2002). If these criteria were not initially met, actions to collapse rating scale categories or deletion of categories would be initiated, in line with the literature (Linacre, 2004).

Evidence of internal structure of the NEQ was then further investigated by monitoring the item goodness-of-fit statistics. WINSTEPS generates both *MnSq* residuals and standardized *z*-values for each of the items of the NEQ. The goodness-of-fit statistics indicate the degree of match between actual responses on the items and expected responses from the Rasch model assertions (Bond and Fox, 2013). Goodness-of-fit was evaluated by infit statistics, as they are viewed as more sensitive to item performance and also more informative when exploring internal scale validity (Wright and Masters, 1982; Bond and Fox, 2013). Furthermore, the *MnSq* fit statistic is preferable for item goodness-of-fit with polytomous data as it is less sensitive to sample size (Smith et al., 2008). The current study chose a sample-size adjusted criterion for item goodness-of-fit set for infit *MnSq* values between 0.7 and 1.3 for the NEQ (Smith et al., 2008). If one or more items would not demonstrate acceptable goodness-of-fit to the model, the items would be removed from the analysis and the iteration process would be repeated until all items met the criterion of acceptable goodness-of-fit.

In order to evaluate the unidimensionality of the NEQ, a principal component analysis of the residuals was also performed (Linacre, 2005). The criterion for unidimensionality was that at least 50% of the total variance should be explained by the first latent variable (Raîche, 2005), and that

no more than 5% should be explained by the largest secondary dimension with an associated eigen value of 2.0, which is an indication of lack of multi-dimensionality.

Evidence of person response validity was then evaluated by monitoring the person goodness-of-fit statistics. The criterion for evaluating person goodness-of-fit was to reject infit *MnSq* values >1.4 associated with a *z*-value >2. It was also accepted that 5% of the sample may fail to demonstrate acceptable goodness-of-fit by chance, without a serious threat to validity (Patomella et al., 2006).

In order to monitor the precision of the converted measures, the person and item separation indices were calculated (Fisher, 1992). The person separation index reflects the number of statistically different strata that the test can identify in the sample, considering the range and precision of the individual person and item estimates. In a similar way, the item separation index reflects the number of statistically different strata that the sample can identify among the items. An index above 1.5 would ensure that the NEQ could differentiate at least two different groups in the sample/among the items.

Finally, a number of Differential Item Functioning (DIF) analyses were performed in order to explore the stability of the response patterns of the NEQ items across sociodemographics, giving further support of validity in relation to internal structure and potential unfairness in testing. This was conducted because it is crucial that an instrument is not biased with regard to any sociodemographics that may otherwise compromise the converted measures, question the validity of the instrument, and influence the interpretation of subsequent findings. The magnitude of DIF was evaluated using the Mantel-Haenszel statistic for polytomous scales using log-odds estimators (Mantel, 1963).

Results

Overall response pattern

Prior to evaluating the categorical responses from the NEQ, all the criteria were met. All rating scale categories were used, which advanced monotonically, and the outfit *MnSq* values for the step category calibrations ranged from 0.89 to 1.11. Only 281 participants out of 564 scored any of the items of the NEQ, and a total of 86% of the person-item data matrix were non-responses, i.e. empty cells (see Table 1). The following item and person validity analyses was thus performed with a limited number of data records, as only 50.9% of the sample reported to have experienced any negative effect of their psychological treatment.

Item goodness-of-fit

The first iteration generating item goodness-of-fit statistics for the 32 items revealed that six items did not meet the criterion for item goodness-of-fit (see Table 2). By removing these items, the next iteration revealed that an additional four items did not meet the criterion and were thus removed. In the third iteration, two more items were removed. Hence, after the third iteration and the removal of 12 items in total (37.5%), the remaining 20 items on the NEQ demonstrated acceptable item fit to the Rasch model assertions. For an overview of the frequencies and average negative impact of each item in the final scale, see Table 3.

Principal component analysis

Following the removal of the twelve items demonstrating misfit, the principal component analysis revealed that the first component explained 62.5% of the total variance, which exceeded the criterion of at least 50% required in order to establish unidimensionality (see Table 2). The second dimension explained an additional 6.3 associated with an eigen value of 3.37, which surpasses the criteria set. By monitoring the item residual loadings, items 15, 11, 3 and 1 loaded more strongly

Table 2. The psychometric properties of the negative effects questionnaire

	NEQ total scale (32 items) (N = 564/281)	NEQ final scale (20 items) (N = 564/264)
Rating scale functioning	All criteria met	All criteria met
Item misfit*		
1st iteration	Item 10, 19, 25, 29, 30, 31	
2nd iteration	Item 5, 8, 9, 21	
3rd iteration	Item 7, 27	
4th iteration		All items met criteria
Variance explained		
1st dimension		62.5%
2nd dimension		6.3%
Person misfit		
N (%)		12 (4.5%)
Maximum score		2 (0.7%)
Minimum score		23 (8.6%)
Person separation index	0.89	1.08
Item separation index	2.01	2.61
Differential item functioning		No differential item functioning

Table 3. Frequencies, means, and standard deviations for the negative effects questionnaire final scale (20 items)[a]

Item	Frequency (%)	M (*SD*)
2: I felt like I was under more stress	106 (37.8)	1.60 (0.86)
13: Unpleasant memories resurfaced	71 (25.3)	1.31 (0.80)
3: I experienced more anxiety	69 (24.6)	1.78 (0.92)
11: I experienced more unpleasant feelings	66 (23.5)	1.50 (0.88)
22: I did not always understand my treatment	55 (19.6)	0.87 (0.82)
26: I felt that the treatment did not produce any results	48 (17.1)	1.75 (1.19)
18: I started thinking that the issue I was seeking help for could not be made any better	44 (15.7)	1.48 (0.88)
1: I had more problems with my sleep	43 (15.3)	1.28 (0.77)
4: I felt more worried	37 (13.2)	1.43 (0.87)
17: I stopped thinking that things could get better	28 (10.0)	1.79 (1.03)
32: I felt that the treatment was not motivating	27 (9.6)	1.89 (1.34)
12: I felt that the issue I was looking for help with got worse	23 (8.2)	1.30 (0.77)
14: I became afraid that other people would find out about my treatment	23 (8.2)	0.87 (0.76)
6: I experienced more hopelessness	23 (8.2)	1.43 (1.08)
24: I did not have confidence in my treatment	21 (7.5)	1.24 (0.89)
16: I started feeling ashamed in front of other people because I was having treatment	15 (5.3)	1.13 (0.99)
20: I think that I have developed a dependency on my treatment	14 (5.0)	0.64 (0.50)
28: I felt that my expectations for the therapist were not fulfilled	10 (3.6)	1.10 (0.74)
23: I did not always understand my therapist	4 (1.4)	1.00 (0.00)
15: I got thoughts that it would be better if I did not exist anymore and that I should take my own life	2 (0.7)	1.50 (0.71)

[a]Based on the number of patients reporting any type of negative effect caused by treatment, N = 281.

on one component, while items 18, 4, 16, 12 and 20, however, loaded more strongly on another (see Table 4).

Person response validity

When evaluating the person response validity, twelve of the 264 participants (4.6%) did not demonstrate acceptable goodness-of-fit to the Rasch model in their responses to the NEQ, which

Table 4. The item residual loadings for the negative effects questionnaire final scale (20 items)

Contrast	Loading	Measure	Infit (MnSQ)	Outfit (MnSQ)	Entry number	Item
1	.77	43.88	1.17	1.11	A 15	15: I got thoughts that it would be better if I did not exist anymore and that I should take my own life
1	.61	38.40	1.03	1.07	B 11	11: I experienced more unpleasant feelings
1	.52	32.35	1.28	1.32	C 3	3: I experienced more anxiety
1	.49	34.85	1.20	1.22	D 1	1: I had more problems with my sleep
2	.26	66.88	.76	.58	E 23	23: I did not always understand my therapist
2	.24	47.55	.89	.98	F 13	13: Unpleasant memories resurfaced
2	.16	36.53	1.03	1.26	G 17	17: I stopped thinking that things could get better
2	.07	57.63	.24	.23	H 28	28: I felt that my expectations for the therapist were not fulfilled
3	−.62	45.89	.90	.96	a 18	18: I started thinking that the issue I was seeking help for could not be made any better
3	−.58	56.32	.86	.90	b 4	4: I felt more worried
3	−.58	54.89	.99	1.27	c 16	16: I started feeling ashamed in front of other people because I was having treatment
3	−.56	55.12	1.12	1.10	d 12	12: I felt that the issue I was looking for help with got worse
3	−.41	70.84	.97	.97	e 20	20: I think that I have developed a dependency on my treatment
3	−.33	70.43	.65	.64	f 14	14: I became afraid that other people would find out about my treatment
3	−.26	54.26	.99	.82	g 6	6: I experienced more hopelessness
2	−.10	63.83	.99	.97	h 22	22: I did not always understand my treatment
2	−.09	41.60	1.24	1.31	i 32	32: I felt that the treatment was not motivating
2	−.05	33.54	1.13	1.14	j 2	2: I felt like I was under more stress
2	−.04	42.36	.96	.90	J 26	26: I felt that the treatment did not produce any results
2	−.01	52.86	.72	.73	I 24	24: I did not have confidence in my treatment

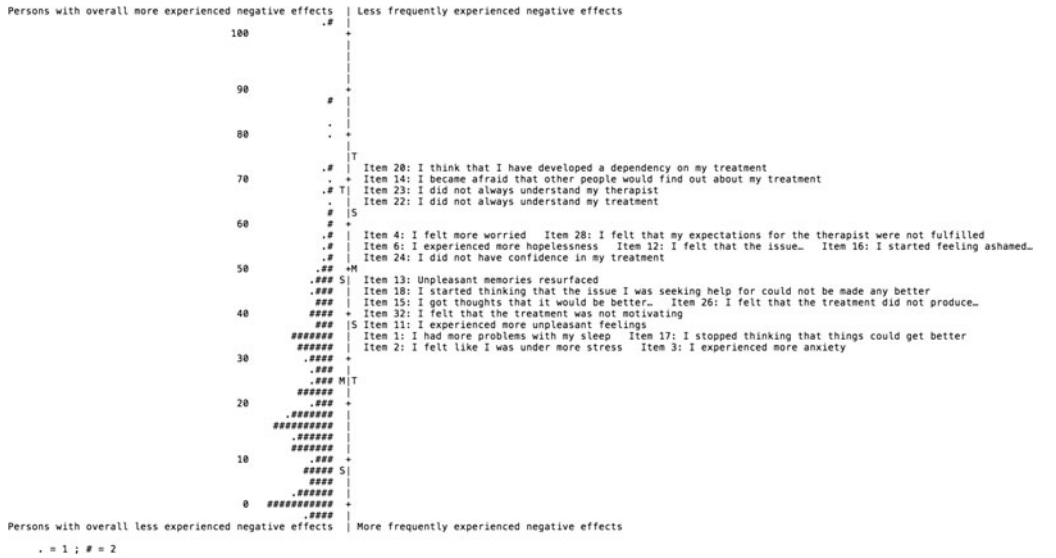

Figure 1. Item-Person map for the negative effects questionnaire final cale (20 items).

met the criterion of up to 5%. Number of participants providing maximum and minimum scores are reported in Table 2.

The person-separation index for the original version of the NEQ, i.e. with 32 items, was 0.89. Moreover, the item-separation index ($N = 281$) was 2.01. After deletion of the 12 NEQ items demonstrating misfit to the Rasch model, the person-separation index increased to 1.08, and the item-separation index scale ($N = 264$) to 2.61 (see Table 2).

The DIF analyses revealed that all of the 20 remaining items of the NEQ functioned in a similar manner across sociodemographics (see Table 2), supporting fairness in testing.

The person-item map is presented in Fig. 1. Items reflecting negative effects more frequently experienced by the sample are placed at the lower end of the continuum, and items reflecting negative effects less frequently experienced by the sample are placed at the higher end of the continuum. In a similar way, participants with fewer experiences of negative effects are placed at the lower end of the continuum, and participants with more experiences of negative effects are placed at the higher end of the continuum.

Discussion

The current study is the first to examine the psychometric properties of an instrument for determining negative effects of psychological treatments using Rasch analysis. In contrast to prior investigations, which have relied on classical test theory (Ladwig et al., 2014; Parker et al., 2013; Rozental et al., 2016), this has enabled an additional investigation of the reliability and validity of persons and items (Waugh and Chapman, 2004), providing a more comprehensive understanding of how negative effects might be assessed. The results suggest that the NEQ exhibits fairness in testing, i.e.it does not demonstrate any bias in terms of the participants' sociodemographics. This important finding suggests that the instrument should yield comparable measures across respondents regardless of gender, age, civil status, educational level, and type of employment, as items are functioning in a similar manner. Also, out of the original 32 items of the NEQ, 12 could be removed as they did not meet the criterion for goodness-of-fit, resulting in a final scale of 20 items that can be downloaded and used for free in clinical and research settings:

www.neqscale.com. Reviewing these items indicate that the factor *failure* is no longer included in the instrument, which may be explained by the fact that it explained less than 3% of the variance in Rozental et al. (2016). From a theoretical perspective, it is also uncertain if *failure* reflects a poor outcome rather than actually experiencing these negative effects during treatment, making it reasonable to exclude the items belonging to this factor from the NEQ. As for the rest of the items that were removed from the instrument, these were primarily related to *dependency* and *quality*, and to a lesser extent *hopelessness* and *symptoms*. Albeit not as clear, it could be argued these items are unrelated to the underlying construct of negative effects or that there is a considerable overlap between them and the items that were retained. For instance, 'I did not have confidence in my treatment' (Item 24) may possibly capture the same concept as 'I felt that my expectations for the treatment were not fulfilled' (Item 27), the latter being excluded. However, it is also important to note that although 12 items did not demonstrate acceptable fit to the Rasch measurement model, indicating that these items demonstrated more unexpected variations in their scores in order to contribute to one underlying measureable construct, they may still add important information about negative effects. Still, it seems that a final scale of 20 items is reliable and could be easier to administer compared with the total scale of 32 items, which should help researchers and therapists to monitor negative effects on a more regular basis.

The rating scale of the instrument also seems to function equally across items, i.e. advancing monotonically, suggesting that the incremental steps of 0 to 4 are appropriate. Several item residuals did, however, load on two components, implying possible multi-dimensionality in the instrument. In relation to the factors obtained by Rozental et al. (2016), the first component is associated with symptoms, while the other is linked to four separate factors. The reason for this finding is unclear and prior research has not discussed the dimensionality or hierarchy of negative effects. Nonetheless, one plausible explanation could be that it reflects a distinction between the subjective experiences of incidents occurring during treatment, e.g. more anxiety, and implications that are interpersonal or social in character, such as dependency and stigma. Strupp and Hadley (1977) considered this issue in their tripartite perspective of psychological treatments, proposing that positive and negative effects might be judged differently by the patient, the therapist, and significant others. Another explanation may be that some negative effects are short term, as in experiencing more unpleasant feelings during treatment, while others are long term, as in believing things cannot improve. This notion has been raised by Castonguay et al. (2010), pointing to the fact that some interventions will never be perceived as particularly pleasant to the patient, even though they are seen as beneficial in the long run. Differentiating those negative effects that are enduring from those that are transient is thus an important research endeavour, preferably by assessing such instances both during treatment and at long-term follow-up. Future studies should also explore if other approaches to examine the instrument (e.g. multi-dimensional Rasch modelling) could yield additional and better solutions to measure negative effects. Still, the findings from the current study indicate that a majority of the items function well enough together to explain a large proportion of the variance and that they also yield acceptable person-fit statistics, which is an important aspect of measurement validity. Given that research on negative effects of psychological treatments is still a fairly new and unexplored territory, psychometric issues such as multi-dimensionality are nevertheless important to consider in order to move the field forward.

As for the rate of negative effects, the number of participants reporting negative effects in the current study was 50.9%, consistent with 58.7% among patients in a psychiatric setting who responded to the INEP (Rheker et al., 2017). However, this number varies significantly between investigations, with rates as high as 92.9% among patients with obsessive-compulsive disorder that were assessed with the Side-effects of Psychotherapy Scale in a study by Moritz et al. (2015), and as low as 5.2% in national survey by Crawford et al. (2016) probing for 'lasting bad effects from the treatment'. Hence different types of assessments and patients will generate different ratios, making it difficult to determine which estimate is more accurate and to compare it across investigations. One of the advantages of implementing Rasch analysis is, however, the possibility to go beyond

just frequencies or levels of symptoms, adjusting for both aspects within a sample. In other words, a person experiencing a large impact on a limited number of items that are rarely perceived among the sample will generate a higher measure of negative effects, compared with a person who is experiencing a moderate impact on a larger number of items that are more often experienced among the sample. Taking this into account, the results from the current study suggest that the instrument has an acceptable person goodness-of-fit, but that it is inappropriate for differentiating distinct subgroups with regard to their experiences of negative effects. This is caused by a relatively large individual standard error associated with each individual measure, as most participants only endorsed a limited number of items (see Table 4). The NEQ is therefore restricted in detecting changes or differences within a specific sample based on their person measures, but is probably suitable for examining differences between clinical trials, settings or interventions by monitoring item difficulty calibrations, e.g. the rate and impact of a particular item, such as placement along the continuum. Further research will help to provide a better estimate of negative effects in psychological treatments by administering the NEQ on a more regular basis during the treatment period, but also to include it in more diverse patient populations.

There are some limitations that need to be addressed when interpreting the results. First, even though the sample was relatively heterogeneous with regard to their sociodemographics, the inclusion and exclusion criteria of each clinical trial may have affected the generalizability of the findings. A majority of the participants were female, middle-aged, in a relationship, having a university degree, and either students or employed, which may have affected the negative effects that were reported. Second, in terms of symptom severity at pre-treatment, the participants were, on average, sub-clinical, at least with regard to the PHQ-9, the GAD-7 and the BBQ, suggesting that they were relatively high functioning. It is possible that another sample would have responded differently on the NEQ, for instance patients with more severe psychiatric disorders than those included in the current study, such as personality disorders, recurrent depression, or eating disorders. Third, given that the participants received psychological treatments mostly administered via the internet or virtual reality, it might be that other formats or theoretical orientations than CBT could result in different negative effects, hence affecting such issues as what items to retain and the principal component analysis. Distributing the NEQ to even more diverse patient populations is thus needed and important to fully understand the occurrence and characteristics of negative effects of psychological treatments. Thus, until further research has been made, some caution is warranted in terms of interpreting the results from the 20 item-version of the NEQ in other formats than the internet, as well as for patients with more severe psychiatric disorders. Fourth, distributing an instrument to patients on a single occasion is problematic, especially concerning incidents that may have been experienced as negative by patients (Rozental et al., 2014). It is possible that the negative effects that were reported were affected by recall bias, primacy-recency effects, and social desirability (Krosnick, 1999), resulting in less valid responses. Future research could therefore include the NEQ on at least one more occasion during the treatment period, for instance at mid-assessment. This should also be accompanied by an investigation of its relationship with outcome, i.e. whether or such incidents affect the long-term benefits.

Author ORCIDs. Alexander Rozental 0000-0002-1019-0245; Johanna Boettcher 0000-0002-8220-9291

Acknowledgements. The authors would like to thank the Swedish Association of Behaviour Therapy (SABT) for a travel grant that allowed A.R. to visit A.K. at the University of Illinois at Chicago to perform the data analysis.

Financial support. This study was made possible in part by a travel grant by the Swedish Association of Behaviour Therapy (SABT). However, the funder had no role in the analyses or drafting of the manuscript.

Conflicts of interest. The authors have no conflicting interests to report.

Ethical statements. The current study adheres to the ethical principles of psychologists and the code of conduct of the American Psychological Association. Ethical approval for the clinical trials was granted at each study location via their respective Regional Ethical Boards, and informed consent was collected from all participants.

References

Andersson, G. (2016). Internet-delivered psychological treatments. *Annual Review of Clinical Psychology*, *12*, 157–179. doi: https://doi.org/10.1146/annurev-clinpsy-021815-093006

Andersson, G., Carlbring, P. and Hadjistavropoulos, H. D. (2017). Internet-based cognitive behavior therapy. In S. G. Hofmann, and G. J. G. Asmundson (eds), *The Science of Cognitive Behavioral Therapy* (pp. 531–549). San Diego, CA: Academic Press.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573. doi: https://doi.org/10.1007/BF02293814

Boettcher, J., Rozental, A., Andersson, G. and Carlbring, P. (2014). Side effects in internet-based interventions for social anxiety disorder. *Internet Interventions*, *1*, 3–11. doi: https://doi.org/10.1016/j.invent.2014.02.002

Bond, T. G. and Fox, C. M. (2013). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. New Jersey, NJ: Psychology Press.

Boswell, J. F., Kraus, D. R., Miller, S. D. and Lambert, M. J. (2015). Implementing routine outcome monitoring in clinical practice: benefits, challenges, and solutions. *Psychotherapy Research*, *25*, 6–19. doi: https://doi.org/10.1080/10503307.2013.817696

Castonguay, L. G., Boswell, J. F., Constantino, M. J., Goldfried, M. R. and Hill, C. E. (2010). Training implications of harmful effects of psychological treatments. *American Psychologist*, *65*, 34–49. doi: https://doi.org/10.1037/a0017330

Crawford, M. J., Thana, L., Farquharson, L., Palmer, L., Hancock, E., Bassett, P., et al. (2016). Patient experience of negative effects of psychological treatment: results of a national survey. *British Journal of Psychiatry*, *208*, 260–265. doi: https://doi.org/10.1192/bjp.bp.114.162628

Fisher, W. (1992). Reliability, separation, strata statistics. *Rasch Measurement Transactions*, *6*, 238. doi: https://doi.org/

Hansen, N. B., Lambert, M. J. and Forman, E. M. (2002). The psychotherapy dose-response effect and its implications for treatment delivery services. *Clinical Psychology: Science and Practice*, *9*, 329–343. doi: https://doi.org/10.1093/clipsy.9.3.329

Hofmann, S. G., Asnaani, A., Vonk, I. J. J., Sawyer, A. T. and Fang, A. (2012). The efficacy of cognitive behavioral therapy: a review of meta-analyses. *Cognitive Therapy and Research*, *36*, 427–440. doi: https://doi.org/10.1007/s10608-012-9476-1

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537–567. doi: https://doi.org/10.1146/annurev.psych.50.1.537

Ladwig, I., Rief, W. and Nestoriuc, Y. (2014). What are the risks and side effects to psychotherapy? – development of an Inventory for the assessment of Negative Effects of Psychotherapy (INEP). *Verhaltenstherapie*, *24*, 252–263. doi: https://doi.org/10.1159/000367928

Lerdal, A., Kottorp, A., Gay, C., Aouizerat, B. E., Lee, K. A. and Miaskowski, C. (2016). A Rasch analysis of assessments of morning and evening fatigue in oncology patients using the Lee Fatigue Scale. *Journal of Pain and Symptom Management*, *51*, 1002–1012. doi: https://doi.org/10.1016/j.jpainsymman.2015.12.331

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*, 85–106.

Linacre, J. M. (2004). Rasch model estimation: further topics. *Journal of Applied Measurement*, *5*, 95–110.

Linacre, J. M. (2005). Rasch dichotomous model vs. one-parameter logistic model. *Rasch Measurement Transactions*, *19*, 1032.

Lindner, P., Frykheden, O., Forsström, D., Andersson, E., Ljótsson, B., Hedman, E., et al. (2016). The Brunnsviken Brief Quality of life scale (BBQ): development and psychometric evaluation. *Cognitive Behaviour Therapy*, *45*, 182–195. doi: https://doi.org/10.1080/16506073.2016.1143526

Loerinc, A. G., Meuret, A. E., Twohig, M. P., Rosenfield, D., Bluett, E. J. and Craske, M. G. (2015). Response rates for CBT for anxiety disorders: need for standardized criteria. *Clinical Psychology Review*, *42*, 72–82. doi: https://doi.org/10.1016/j.cpr.2015.08.004

Lovibond, P. F. and Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*, 335–343. doi: https://doi.org/10.1016/0005-7967(94)00075-U

Löwe, B., Kroenke, K., Herzog, W. and Gräfe, K. (2004). Measuring depression outcome with a brief self-report instrument: sensitivity to change of the Patient Health Questionnaire (PHQ-9). *Journal of Affective Disorders*, *81*, 61–66. doi: https://doi.org/10.1016/S0165-0327(03)00198-8

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, *58*, 690–700. doi: https://doi.org/10.2307/2282717

McHugh, R. K. and Barlow, D. H. (2010). The dissemination and implementation of evidence-based psychological treatments: a review of current efforts. *American Psychologist*, *65*, 73–84. doi: https://doi.org/10.1037/a0018121

Mechler, J. and Holmqvist, R. (2016). Deteriorated and unchanged patients in psychological treatment in Swedish primary care and psychiatry. *Nordic Journal of Psychiatry*, *70*, 16–23. doi: https://doi.org/10.3109/08039488.2015.1028438

Moritz, S., Fieker, M., Hottenrott, B., Seeralan, T., Cludius, B., Kolbeck, K., et al. (2015). No pain, no gain? Adverse effects of psychotherapy in obsessive-compulsive disorder and its relationship to treatment gains. *Journal of Obsessive-Compulsive and Related Disorders*, *5*, 61–66. doi: https://doi.org/10.1016/j.jocrd.2015.02.002

Muris, P. and Merckelbach, H. (1996). A comparison of two spider fear questionnaires. *Journal of Behavior Therapy and Experimental Psychiatry*, *27*, 241–244. doi: https://doi.org/10.1016/S0005-7916(96)00022-5

Parker, G., Fletcher, K., Berk, M. and Paterson, A. (2013). Development of a measure quantifying adverse psychotherapeutic ingredients: the Experiences of Therapy Questionnaire (ETQ). *Psychiatry Research*, *206*, 293–301. doi: https://doi.org/10.1016/J.Psychres.2012.11.026

Patomella, A-H., Tham, K. and Kottorp, A. (2006). P-drive: assessment of driving performance after stroke. *Journal of Rehabilitation Medicine*, *38*, 273–279. doi: https://doi.org/10.1080/16501970600632594

Raîche, G. (2005). Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions*, *19*, 1012.

Rheker, J., Beisel, S., Kräling, S. and Rief, W. (2017). Rate and predictors of negative effects of psychotherapy in psychiatric and psychosomatic inpatients. *Psychiatry Research*, *254*, 143–150. doi: https://doi.org/10.1016/j.psychres.2017.04.042

Rozental, A., Andersson, G., Boettcher, J., Ebert, D. D., Cuijpers, P., Knaevelsrud, C., et al. (2014). Consensus statement on defining and measuring negative effects of Internet interventions. *Internet Interventions*, *1*, 12–19. doi: https://doi.org/10.1016/j.invent.2014.02.001

Rozental, A., Boettcher, J., Andersson, G., Schmidt, B. and Carlbring, P. (2015). Negative effects of Internet interventions: a qualitative content analysis of patients' experiences with treatments delivered online. *Cognitive Behaviour Therapy*, *44*, 223–236. doi: https://doi.org/10.1080/16506073.2015.1008033

Rozental, A., Kottorp, A., Boettcher, J., Andersson, G. and Carlbring, P. (2016). Negative effects of psychological treatments: an exploratory factor analysis of the Negative Effects Questionnaire for monitoring and reporting adverse and unwanted events. *PLoS One*, *11*, e0157503. doi: https://doi.org/10.1371/journal.pone.0157503

Rozental, A., Magnusson, K., Boettcher, J., Andersson, G. and Carlbring, P. (2017). For better or worse: an individual patient data meta-analysis of deterioration among participants receiving internet-based cognitive behavior therapy. *Journal of Consulting and Clinical Psychology*, *85*, 160–177. doi: https://doi.org/10.1037/ccp0000158

Shea, T. L., Tennant, A. and Pallant, J. F. (2009). Rasch model analysis of the Depression, Anxiety and Stress Scales (DASS). *BMC Psychiatry*, *9*, 21. doi: https://doi.org/10.1186/1471-244X-9-21

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G. and Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology*, *8*(33). doi: https://doi.org/10.1186/1471-2288-8-33

Spector, P. E. (1992). *Summated Rating Scale Construction: An Introduction*. Newbury Park, CA: Sage.

Spitzer, R. L., Kroenke, K., Williams, J. B. W. and Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder. *Archives of Internal Medicine*, *166*, 1092. doi: https://doi.org/

Strupp, H. H. and Hadley, S. W. (1977). A tripartite model of mental health and therapeutic outcomes. With special reference to negative effects in psychotherapy. *The American Psychologist*, *32*, 187–196. doi: https://doi.org/10.1037/0003-066X.32.3.187

Suh, C. S., Strupp, H. H. and O'Malley, S. S. (1986). The Vanderbilt process measures: the Psychotherapy Process Scale (VPPS) and the Negative Indicators Scale (VNIS). In L. S. Greenberg (ed), *The Psychotherapeutic Process: A Research Handbook* (pp. 285–323). New York, NY, USA: Guilford Press.

Vlaescu, G., Alasjö, A., Miloff, A., Carlbring, P. and Andersson, G. (2016). Features and functionality of the Iterapi platform for internet-based psychological treatment. *Internet Interventions*, *6*, 107–114. doi: https://doi.org/10.1016/j.invent.2016.09.006

Waugh, R. and Chapman, E. (2004). An analysis of dimensionality using factor analysis (true-score theory) and Rasch measurement: what is the difference? Which method is better? *Journal of Applied Measurement*, *6*, 80–99.

Wright, B. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*, 97–116. doi: https://doi.org/10.1111/j.1745-3984.1977.tb00031.x

Wright, B. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*, 3–24. doi: https://doi.org/10.1080/10705519609540026

Wright, B. and Masters, G. N. (1982). *Rating Scale Analysis: Rasch Measurement*. Chicago, IL: MESA Press.

Cite this article: Rozental A, Kottorp A, Forsström D, Månsson K, Boettcher J, Andersson G, Furmark T, and Carlbring P (2019). The Negative Effects Questionnaire: psychometric properties of an instrument for assessing negative effects in psychological treatments. *Behavioural and Cognitive Psychotherapy* **47**, 559–572. https://doi.org/10.1017/S1352465819000018