V2, 03/2017

The Hamburg Burnout Inventory (HBI)

Background and Some Early Results

Matthias Burisch[1]

University of Hamburg

burisch@uni-hamburg.de

1. Overview

This technical report will briefly review the history of the *Hamburg Burnout Inventory (Hamburger Burnout-Inventar; HBI)*, an instrument which was developed out of frustration over the prevalent burnout questionnaires, the *Maslach Burnout Inventory (MBI;* Maslach & Jackson, 1986; Maslach, Jackson, & Leiter, 1996) and the *Tedium Measure (TM;* Aronson, Pines & Kafry, 1983), later rechristened as the *Burnout Measure.* Although these two instruments still enjoy the status of a „gold standard" and dominate more than 90% of published empirical burnout research (Schaufeli & Enzmann, 1998, p. 71; Rösing, 2003, p. 69-75) their shortcomings are quite obvious. To mention only some, the MBI's validity is questionable at best (Burisch, 1984b, unpublished) while the TM is both highly reliable and valid, but even more undifferentiated than the MBI, and probably a misnomer (Enzmann, 1996). Moreover, the original version of the MBI was applicable only to people in service jobs, since many of the items referred to „recipients" of those services, something which has in the meantime been corrected by the publication of a „general" version (Maslach, Jackson, & Leiter, 1996).

The HBI measures facets of burnout in ten very short scales and one additional item. It has been shown to possess adequate reliability and validity vis-á-vis peer ratings.

[1] Thanks are due Bianka Giesa, Felix Frühauf and Kirsten Steinhoff, who worked on the HBI in its early stages, and Maren Hagge, Manuela Mielke, Anna-Maria Tolke, Friedrich von Herder, and Uli Weber, who later ran projects on the instrument or contributed data. Erich Hotter generously supplied several very large samples. Uwe Schaarschmidt let us use his AVEM questionnaire and provided software to score it free of charge. I am also grateful to Alexander Harbaugh for editing my raw translation into English (and for much more!) and Catherine Vasey and Alexander Harbaugh for preparing the French version. Rainer Kurz later polished the English version. Thanks are also due Frank T. Petermann who ran *SwissBurnout* while it was operational, which made the HBI known worldwide. Lew Goldberg sacrificed precious beach time in order to make this manuscript more readable; his contribution (and much more!) is highly appreciated.

Recently, it has gained enormous popularity due to its availability in a self-scoring and instant-feedback version on the website of *SwissBurnout.ch* (SWB), a Swiss nonprofit organization. Between early 2006 and June 2007 about 40.000 individuals had left their (anonymous) traces there; in 2013, when the service was discontinued, that number had risen to over 300.000.

This report will give a brief background to the instrument and draw some tentative substantive conclusions, particularly from the second of two batches of SWB online data.

2. History of the Instrument

The *Hamburg Burnout Inventory (Hamburger Burnout-Inventar; HBI)* was developed during the late eighties and early nineties of the last century in collaboration with several graduate students. Originally, it comprised about 30 constructs and more than 200 items. Gradually, with a mixture of the deductive and the inductive approach to scale development (Burisch, 1984a), i.e. with the application of good common sense and information from incoming empirical data, the scope narrowed. Many of the original constructs proved to be indistinguishable, others collapsed under the evidence from item analyses. What remained after the initial stage — not very well documented since two of the thesis projects were never finished — was 39 items in 10 scales. A little later, item 40 („I am in a crisis right now and cannot find a way out“) was added. It serves as an 1-item capsule measurement of burnout and is not scored for any of the scales.

Compared to standard personality inventories, both German language and English, an average scale length of just 3.9 items may appear daringly short. However, the well-known relationship between scale lenght and reliability does not hold for scale length and *validity* in the typical case. To obtain maximally valid scales, three to eight items may be optimal (Burisch, 1997).

This 40-item questionnaire has not been changed in the meantime. An attempt to augment it with a few items and two additional scales, *Loss of Meaning* and *Engagement,* (Tolke, 2013) failed to bring in encouraging results.

The HBI has been employed in several theses, dissertations, and other projects so far. To my knowledge, there are few published accounts of this is, e.g. Frick & Filipp (1997) or Wurm et al. (2016).

In the spring of 2006, the HBI was posted on *SwissBurnout*‘s website; soon after, it was made self-scoring; i.e., respondents received immediate feedback of test results in the form of a profile.Between March and November of 2006 more than 17.000 users had

completed the German language HBI on the *SwissBurnout* website. These data, omitting incomplete records, comprise SWB Sample 1 (N = 16.273).

Late in 2006, some technical improvements were implemented. An English and a French version were added. Feedback was provided only if the respondent had answered *all* items. Moreover, more differentiated demographics were introduced. Records turned in between Dec 28, 2006 and April 23, 2007 comprise SWB Sample 2 (N = 15.939); more details on SWB2 below. In 2013 the HBI service on the SWB website was discontinued.

Since early 2011 the HBI is available for instant scoring and extensive feedback of results, for a fee, on www.burnout-institut.eu, the website of the Burnout-Institut Norddeutschland (BIND). A short form of the inventory (HBI21) can be found on http://www.cconsult.info/selbsttest/burnout-test.html; minimal feedback of results is offered free of charge.

The instrument‘s major advantages to this date seem to be threefold: (a) With ten scales and an additional item, it provides a much more differentiated picture than the MBI or the TM, although admittedly many of the scales intercorrelate substantially. (b) The available validity information is encouraging. (c) There are rough norms for the German version, whereas those are lacking for both the MBI and the TM.

3. Technical Information

Scale names and number of items are presented in Tab. 1 below. Internal consistency of the HBI scales can be estimated on the basis of eleven samples, retest validity was assessed in one sample. Validity was investigated in four studies. In the sections that follow, data sources are briefly characterized in chronological order.

3. 1 Reliability

(a) Frühauf (1990) administered an early HBI version to a total of 313 adults, including 75 students, many of whom worked part-time as well, and 238 working men and women. Those volunteers, recruited by student assistents, also filled in the *Tedium Measure* and the *Freiburger Persönlichkeits-Inventar* (FPI-R; Fahrenberg, Selg, & Hampel, 1984) for comparison. The first 39 items of the present HBI version were selected on the basis of that study.

(b) Steinhoff (1991; unpubl.) had a sample of 182 adults fill in the 40-item HBI version (CRISIS item added in the meantime) and also name two peers who would independently

and anonymously rate her or him directly on 9-point rating scales. She also administered the short form of the *Freiburger Persönlichkeits-Inventar* (FPI-K; Fahrenberg & Selg, 1970) which contains eight 7-item scales. Subjects were again recruited via student assistents. Care was taken to avoid mentioning „burnout“ to minimize selection effects.

(c) The study by Hagge (2005) followed the same outline as Steinhoff‘s with N = 77 subjects. For benchmarking purposes, she applied the *NEO-FFI* (Borkenau & Ostendorf, 1993) and the *Oldenburg Burnout Inventory* (Ebbinghaus, 1986; Demerouti 1999), which comprises two scales, DISTANCE and EXHAUSTION, each eight items long.

(d, e) Samples SWB1 (N = 16.273; 45% female; mean age 39.5 yr.) and SWB2 (German language subsample only; N = 14.123; 45% female; mean age 40.6 yr.) were described above; more on SWB2 follows. Those data were collected in 2006 and 2007, respectively. A *caveat* is in order here: There is no guarantee that these samples do not contain multiple entries from individuals who „took their pulse“ on a weekly basis, say. Nor can we exclude the possibility of people wo just wanted to „play“ with the device, curious to know just how high or low anyone can get by clicking appropriate buttons. This is why an additional norming system, based on sample SWB1, was withdrawn after a little while. For more details, see section *4. Substantive Findings.*

(f) Hotter (2009, unpubl.) obtained HBI protocols online from 774 Austrian judges (51% female, mean age = 43 yr.). Study participants followed an invitation published by professional associations.

(g) von Herder (2011) administered the HBI and the AVEM (Schaarschmidt & Fischer 2008) to a sample of 47 midlevel managers (18% female; mean age 42 yr.) from one South German medium-size engineering company. Of these, 33 (70%) also took part in a retest about eight weeks later. Retest correlations were computed from this subsample, whereas internal consistency estimates are based on the first testing and all 47 participants.

(h) Tolke (2013) employed the same basic procedure as Steinhoff (1991) and Hagge (2005); see above. This turned in 70 HBI protocols (47% response rate) and two sets of peer ratings each. Her sample excluded students and subjects under 18 years of age (mean age = 39); 69% female. She also investigated a couple of fresh questionnaire items which were not followed up further, however. Participants who volunteered to provide their postal addresses received written feedback of their test results.

(i) Hotter (2014, unpubl.) collected an sample of 6.249 Austrian school teachers (48% female) who had followed an invitation from their teachers‘ union to submit data online.

(j) Weber (2014) used the HBI as one of several outcome variables to test the effects of his own online burnout-prevention program. Only the 861 pretest HBI protocols (69% female; mean age 47 yr.) were used for the present analyses.

(k) Mielke (2014) had prospective outpatients at a large treatment center in Hamburg take the HBI immediately prior to their initial interview. Sixteen of these patients (63% female; mean age 44 yr.) were later rated by one clinician each on corresponding 9-point rating scales. Because of the tiny sample size these data were only used to calculate validity, not reliability, estimates.

(l) Hotter (2016, unpubl.) obtained another sample of 10.674 HBI protocols from Austrian teachers (81% female; mean age 47 yr.) in the same manner as for sample (i).

Table 1
Scales, Scale Lengths, and Reliability Estimates for the Hamburg Burnout Inventory

SAMPLE		**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	**l**	**wM**	**Retest r_{tt}**
SAMPLE N	**m**	**313**	**182**	**77**	**16273**	**14123**	**774**	**47**	**70**	**6249**	**861**	**10674**		**33**
EE	5	83	87	86	91	91	93	92	88	94	91	93	**92**	87
PA	3	74	63	67	82	71	77	63	30/54	69	77	68	**74**	72
DIST	4	71	68	65	76	75	71	61	59	72	77	71	**74**	79
DEP	3	70	62	77	68	70	74	64	68	77	77	76	**72**	80
HELPL	4	83	73	71	88	87	88	87	84	87	89	86	**87**	79
VOID	4	80	74	73	88	87	88	81	79	88	88	87	**87**	84
TD	5	87	89	88	89	91	91	87	93	92	93	91	**91**	73
INUN	3	71	71	73	84	85	85	89	78	86	84	85	**85**	89
OTAX	5	80	72	81	87	85	88	80	82	86	87	84	**86**	70
AGG	3	75	56	74	80	79	80	63	74	78	79	78	**79**	77
Mean	**3.9**	**77**	**72**	**76**	**83**	**82**	**84**	**77**	**74**	**83**	**84**	**82**	**82**	**80**

Legend: *m* = scale length. For information on samples *a* through *l,* see above. *wM* = weighted mean of alphas in columns *a* through *l.* Scale Names: *EE* = EMOTIONAL EXHAUSTION. *PA* = PERSONAL ACCOMPLISHMENT (reversed). *DIST* = DISTANCE. *DEP* = DEPRESSIVE REACTION TO STRESS. *HELPL* = HELPLESSNESS. *VOID* = INNER VOID. *TD* = TEDIUM. *INUN* = INABILITY TO UNWIND. *OTAX* = OBERTAXING ONESELF. *AGG* = AGGRESSIVE REACTION TO STRESS. Decimal points omitted throughout.

For easy reference, Table 1 summarizes all available figures for the HBI‘s reliability. Cronbach’s coefficient alpha as an estimate of internal consistency has come under criticism for some time because it underestimates the true value systematically. Sijtsma (2009) recommends a coefficient known as the Greatest Lower Bound (glb) instead, but software to compute it is not easily accessible, and glb has been shown to be misleading

for sample sizes under 1000. As a substitute, the same author suggests using Guttman's coefficient λ_2. This was done for one of the largest subsample, Hotter's 10674 teachers, see sample (l) above. Since in no case λ_2 exceeded alpha by more than .006, alpha will be reported throughout.

As can be seen in Tab. 1, reliability estimates from smaller samples tend to be lower than those from larger ones. In fact, alpha for scale PERSONAL ACCOMPLISHMENT in sample h was only .30 (which increased to .54 when one outlier subject was deleted). Idiosyncrasies like these, probably due to erratic responding, exert a stronger influence when sampe size is modest.

To simplify the picture, the second column from the right of Tab. 1 *(wM)* contains the weighted means of the eleven sample coefficients. These range from .72 (for DEPRESSIVE REACTION) to .92 (for EMOTIONAL EXHAUSTION). The grand mean of the ten scale alphas is .82.

As mentioned above, retest reliability has been studied by von Herder (2011; sample g, N = 33). Test-retest correlations for this very small sample are also included, in the rightmost column of Tab. 1.

3.2 Validity

Table 2 contains validity information from four separate studies, namely those by Steinhoff (1991, sample b; N = 182), Hagge (2005, sample c; N = 77), Tolke (2013, sample h; N = 70), and Mielke (2014, sample k; N = 16). For the first three of these (samples b, c, and h), the results blocks contain in their leftmost column r_k, a coefficient of interrater agreement, which is simultaneously an estimation of the average rating's reliability (Winer, 1962, p. 124). The correlations between scale scores (self ratings) and averaged peer ratings, r_{tc}, are next. These represent what is normally reported as validity coefficients. The rightmost column of each block give those coefficients divided by the square root of r_k (correction for attenuation). The latter coefficients, $r_{tc\text{-}korr}$, estimate the validity coefficients that would have resulted had the criterion (the averaged peer ratings) been measured error-free.

No correction is possible for the validities in the last column of Table 2, because in that case only one rating was obtained, whereas in the first three studies there were two raters for each subject.

3.3 Discussion

As can be gleaned from Table 1, there is considerable — though far from perfect agreement among the *reliability* estimates from the various samples. Across the samples,

EMOTIONAL EXHAUSTION (mean r_{tt} = .92) and TEDIUM (mean r_{tt} = .91) lead the field, while DEPRESSIVE REACTION (.72), DISTANCE (.74), and PERSONAL ACCOMPLISHMENT (.74) trail it.

As was already noted, the larger samples (i.e. d, e, f, i, j, and l) tend to yield higher reliability estimates than the smaller samples (i.e. a, b, c, g, and h) do. In fact, the mean alphas for the former range from .82 to .84, while those from the latter range from .72 to .77.

Table 2
Validity Estimates for the Hamburg Burnout Inventory

SAMPLE			**b**			**c**				**h**			**k**
	m	**r_k**	**r_{tc}**	**$r_{tc\text{-}corr}$**		**r_k**	**r_{tc}**	**$r_{tc\text{-}corr}$**		**r_k**	**r_{tc}**	**$r_{tc\text{-}corr}$**		**r_{tc}**
EE	5	61	46	59		60	39	50		37	36	59		49
PA	3	54	30	41		35	27	46		64	25	31		45
DIST	4	60	42	54		74	47	55		62	38	48		40
DEP	3	63	35	44		68	23	28		59	50	65		46
HELPL	4	55	40	53		64	39	49		51	51	71		34
VOID	4	65	22	27		69	09	11		37	32	53		58
TD	5	67	46	56		73	33	39		26	39	76		26
INUN	3	61	23	30		68	39	47		55	28	38		39
OTAX	5	61	38	49		50	28	40		63	28	35		33
AGG	3	50	28	40		29	17	32		40	30	47		50
Mean	**3.9**	**60**	**35**	**46**		**61**	**30**	**40**		**46**	**36**	**54**		**42**

Legend: *m* = scale length. For information on samples *b* through *k,* see Section 3.1. r_k = coefficient of interrater agreement. r_{tc} = validity coefficient. $r_{tc\text{-}corr}$ = validity coefficient corrected for attenuation. Scale Names: *EE* = EMOTIONAL EXHAUSTION. *PA* = PERSONAL ACCOMPLISHMENT (reversed). *DIST* = DISTANCE. *DEP* = DEPRESSIVE REACTION TO STRESS. *HELPL* = HELPLESSNESS. *VOID* = INNER VOID. *TD* = TEDIUM. *INUN* = INABILITY TO UNWIND. *OTAX* = OVERTAXING ONESELF. *AGG* = AGGRESSIVE REACTION TO STRESS. Decimal points omitted throughout.

In view of the shortness of the scales — mean length 3.9 items — the reliability figures are remarkably high. This is also true for retest reliability, presented in the table's last column.

The *validities* in Table 2 range from quite good to downright unsatisfactory. There is a coefficient of .09 for VOID in the Hagge sample (c), which rises to .22 , however, when partialling gender out. The Hagge (2005) study in general delivered the lowest coefficients, with a mean (uncorrected) validity of .30 vis-a-vis corresponding figures of .35 (Steinhoff 1991), .36 (Tolke 2013), and .42 (Mielke 2014). Admittedly, the Mielke correlations have a very wide confidence interval. However, they represent the first application of the HBI in a clinical setting.

To put these results into perspective, the mean validity coefficient of the *FPI-K,* an established though now obsolete inventory containing eight scales of seven items each, was only .29 in the Steinhoff sample, where the HBI made it to .35. The HBI fared somewhat worse in Hagge's study where the comparison was with the *NEO-FFI* (five scales of twelve items each) and the *OLBI* (two scales of eight items each). While the *NEO-FFI* scored an average validity coefficient of .40 (corrected .45) and the *OLBI* of .38 (corrected .47), the HBI only reached .30 (corrected .40).

Secondly, remember that many HBI constructs are probably hard to observe and rate from the outside. Thus, although the r_k value for VOID in the above example is the third highest at .69, reflecting good agreement among raters, these ratings may be of limited value. „Inner Void" may be an experience people are wary to communicate even to their closest associates because of its ego-threatening impact. — Admittedly, this is speculative and after the fact.

4. Some Substantive Findings

4.1 Description of SWB Sample 2

Respondents in SWB Sample 2 (with few exceptions) provided the following demographic information: gender, year of birth, nationality (in 13 categories), profession (to be typed in), occupational status (6 categories), and workplace (i. e., Swiss canton vs. „outside Switzerland"; the latter category making up 71%). The language used (German = DE, English = EN, French = FR) and the access date were stored automatically.

A breakdown of the sample by language, gender, and occupational status is contained in Table 3.

Table 3
Gender and Occupational Status for Language Subgroups in Sample SWB2

		GERMAN			ENGLISH			FRENCH	
Status	**male**	**female**	**% of DE**	**male**	**female**	**% of EN**	**male**	**female**	**% of FR**
empl. with exec. functions	1000 (13%)	441 (7%)	**10**	80 (18%)	39 (14%)	**16**	71 (13%)	43 (8%)	**11**
empl. with line resp.	2386 (31%)	1413 (22%)	**27**	149 (33%)	67 (24%)	**30**	151 (28%)	82 (15%)	**21**
empl. without line resp.	2578 (33%)	2808 (44%)	**38**	76 (17%)	45 (16%)	**17**	192 (36%)	290 (53%)	**44**
independent	883 (11%)	519 (8%)	**10**	58 (13%)	39 (14%)	**13**	55 (10%)	35 (6%)	**8**
other	710 (9%)	913 (14%)	**12**	61 (14%)	65 (23w%)	**17**	58 (11%)	76 (14%)	**12**
without employment	172 (2%)	300 (5%)	**3**	27 (6%)	23 (8%)	**7**	11 (2%)	23 (4%)	**3**
Sum	**7729 (100%)**	**6394 (100%)**		**451 (100%)**	**278 (100%)**		**538 (100%)**	**549 (100%)**	
% of Language	**55**	**45**	**100**	**62**	**38**	**100**	**50**	**50**	**100**

The three language-specific subsamples do not differ very much as to gender composition; however, the FR subsample is practically balanced, whereas speakers of DE are more likely (55%) to be male and the EN subsample is predominantly (62%) male. Professional status seems to be quite similar for DE and FR, whereas the EN subsample stands somewhat out. No less than 46% of it is in the upper echelons of employees with executive functions or line responsibility (DE: 37%; FR 32%) and 13% of it is independent (DE: 10%; FR: 8%). At the same time, 7% of the EN subsample are unemployed, compared to only 3% of the DE and FR subsamples, respectively. The average age (computed simply as the difference between 2007 and year of birth) is lowest for EN (33.8) and highest for DE (40.6); the French mean is 38.5.Thus, the EN subsample can be expected to score somewhat differently in terms of burnout. Responses came literally from all over the world, as Table 4 shows, with Germany, Switzerland, and Austria contributing most.

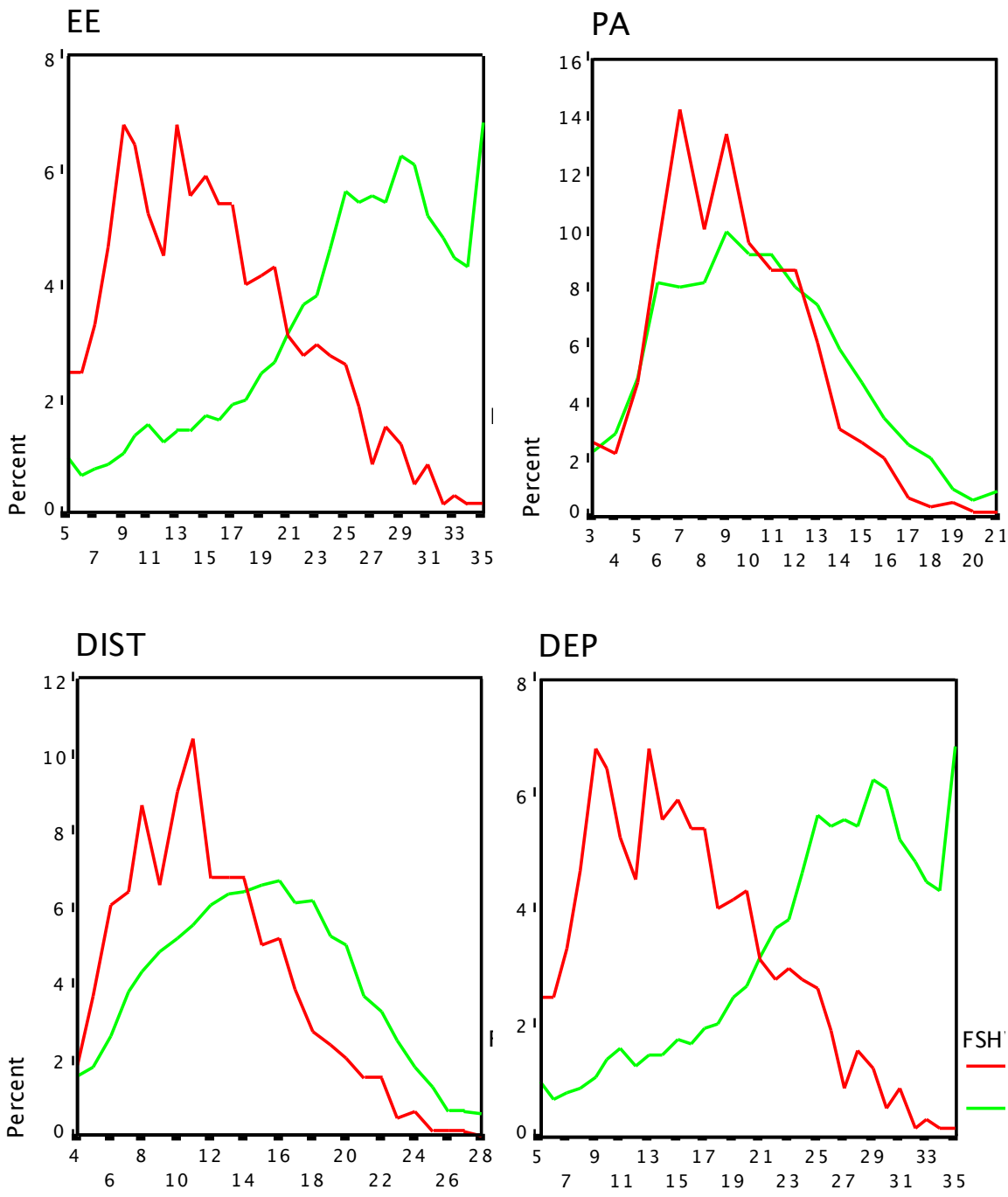
Table 4
SWB2 in Terms of Geography

Asia and Oceania	116
Austria	1046
Eastern Europe	76
England	108
France	324
Germany	9511
Italy	119
Latin America	48
Mediterranean countries	164
North America	373
Other countries the EEC	310
Sub-Saharan Africa	35
Switzerland	3703
missing	6
Total	**15939**

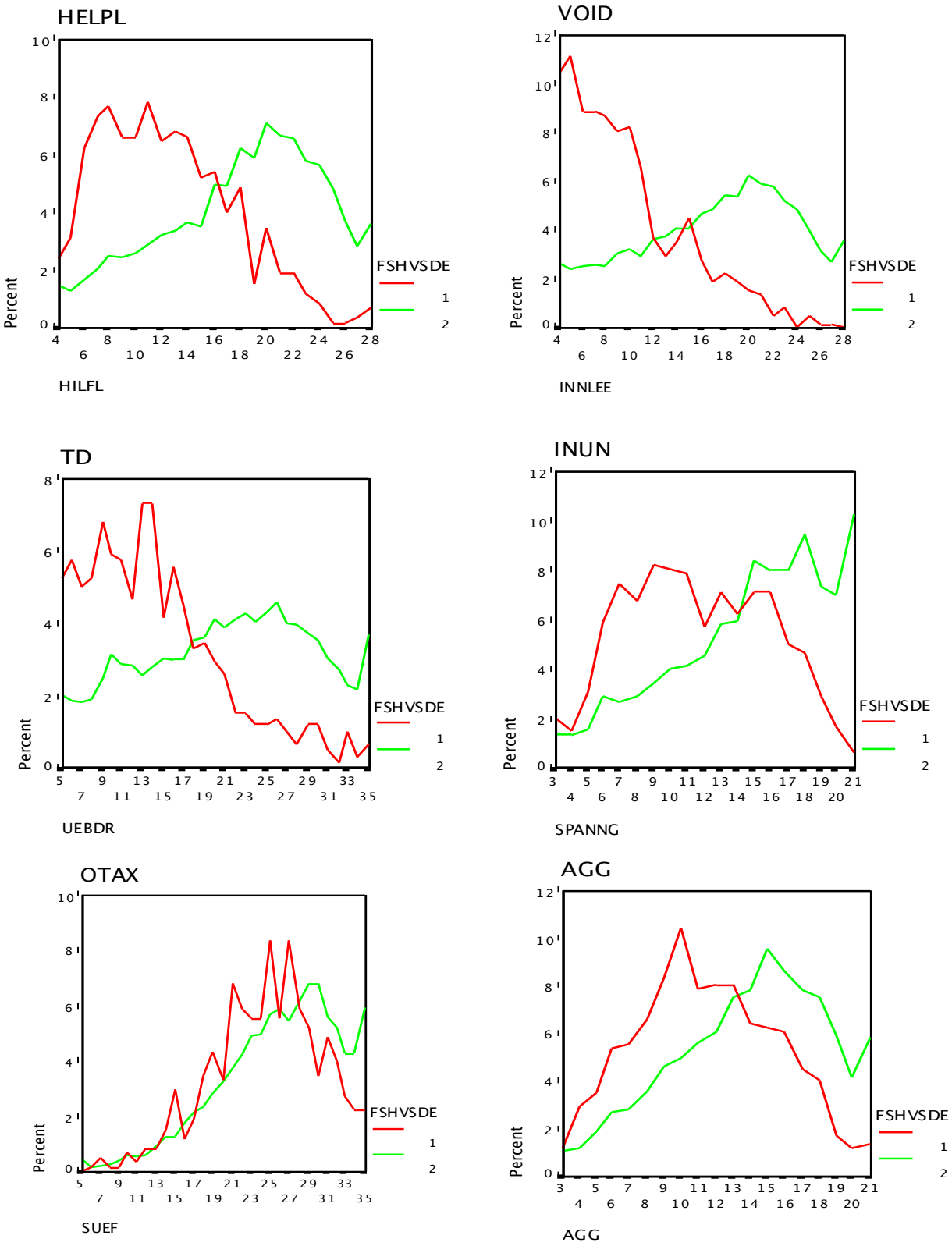
4.2 Comparing Online Samples with Paper-and-Pencil Samples

Probably the most striking finding from SWB samples 1 and 2 is that quite persistently those online samples scored dramatically higher on burnout scales than did earlier samples, using paper versions of the HBI (see Figure 1).

Figure 1
Comparing Samples Frühauf, Steinhoff, & Hagge with Sample SWB 2 (German only) on All HBI scales



Legend: Red curve = combined sample Frühauf, Steinhoff & Hagge (N = 572; all paper-pencil). Green curve: Sample SWB 2, DE only (N = 14123, all online).



Legend: Red curve = combined sample Frühauf, Steinhoff & Hagge (N = 572; all paper-pencil). Green curve: Sample SWB 2, DE only (N = 14123, all online).

Note also the more or less marked upward spike at the right hand end of the green curves above. This is an example of a so-called ceiling effect in the scales. It means that a sizable proportion of SWB's online HBI users would have been prepared to describe their lot in even more negative ways, had test items been provided to do so.

Why this? From hindsight, this may seem unsurprising. After all, users of *SwissBurnout*'s website, even if they got there inadvertently (i.e., not by having typed *burnout* into *Google*), must have possessed enough curiosity and invested the ten minutes to fill in the questionnaire. It is plausible that most of them did *not* do so out of sheer curiosity. In contrast, participants in the studies by Frühauf, Steinhoff, and Hagge (see section 3.1 above) were approached by research assistants who had been warned not to mention the „B word", in order not to wake any sleeping dogs. Thus, although the representativeness of these early samples may be questioned — they were acquaintances or acquaintances of acquaintances of psychology students living in Northern Germany, and students were overrepresented as subjects — they most probably came closer to the „statistical norm" in Germany. At any rate, they were not systematically preselected in terms of burnout.

Before we settle for this explanation, let us examine two competing ones. One is that the HBI item content might somehow have „aged" between the early nineties and 2007. (This would more easily explain a *decline* in burnout scores, though.) Another is that living and working conditions in pertinent areas have sufficiently deteriorated over the past years to be reflected in rising scores.

Fortunately, the data collected by Hagge in 2004 and 2005 are at hand to throw some light on these alternative explanations. Hagge's (2005) German only sample of 77 adults (53% female; mean age = 39.7 yr.) was biased in terms of education: No less that 68% had graduated from high school („Abitur") and 39% even held some academic degree. However, no students were included and socio-economic composition resembled that of SWB sample 2 (where pertinent information is available) more closely than the early samples. Does this contemporary sample exhibit signs of *more* burnout than in the early samples?

As Table 5 shows, the answer is a quite unequivocal *no.* There were only three significant differences, and two of these pointed out *less* burnout in the 2000s decade than in the former. There are some mean differences to the contrary, namely *more* EMOTIONAL EXHAUSTION, *more* INNER VOID, *more* INABILITY TO UNWIND, *more* OVERTAXING ONESELF in the 2000s. But there are also differences in the opposite direction: *Less* DISTANCING, *less* HELPLESSNESS, *less* TEDIUM. And all of them are insignificant, i.e., too small not to be explained by chance (sampling error).

Table 5
Comparing Early Samples Frühauf & Steinhoff with 2005 Sample Hagge

	Samples Frühauf & Steinhoff		Sample Hagge		
	Mean	**SD**	**Mean**	**SD**	***p***
01 Emotional Exhaustion	15.4	6.4	16.3	6.6	n.s.
02 Personal Accomplishment (rev.)	9.0	3.2	10.9	2.7	.00
03 Distance	11.9	4.6	11.4	4.6	n.s.
04 Depressive Reaction to Stress	11.0	3.8	9.7	3.6	.01
05 Helplessness	12.5	5.2	12.0	4.4	n.s.
06 Inner Void	9.8	5.0	9.9	4.6	n.s.
07 Tedium	14.5	6.9	12.9	6.0	n.s.
08 Inability to Unwind	11.5	4.4	12.3	3.9	n.s.
09 Overtaxing	24.3	5.8	25.4	5.8	n.s.
10 Aggressive Reaction to Stress	11.6	4.2	10.5	3.9	.05

Legend: Combined sample Frühauf & Steinhoff (N = 495; all paper-pencil). Sample Hagge, (N = 77; all paper-pencil). *p* = type I error probability (two-sided).

How about SWB sample 1(all German)? Frequency distributions (not shown here) are strikingly similar to the SWB 2 German subsample, with the exception of having slightly smaller means in all cases. (This latter effect is consistently significant — to be expected with samples that large — but not very strong.)

We may thus conclude with some level of confidence that the dramatically high burnout scores of SWB website users reflect a „real“ self-selection phenomenon: People who visited the website suspected they were in a burnout process. To verify this, they took the test. A certain part of them got confirmation of their hunch.

4.3 Comparing the German, English, and French subsamples of SWB2

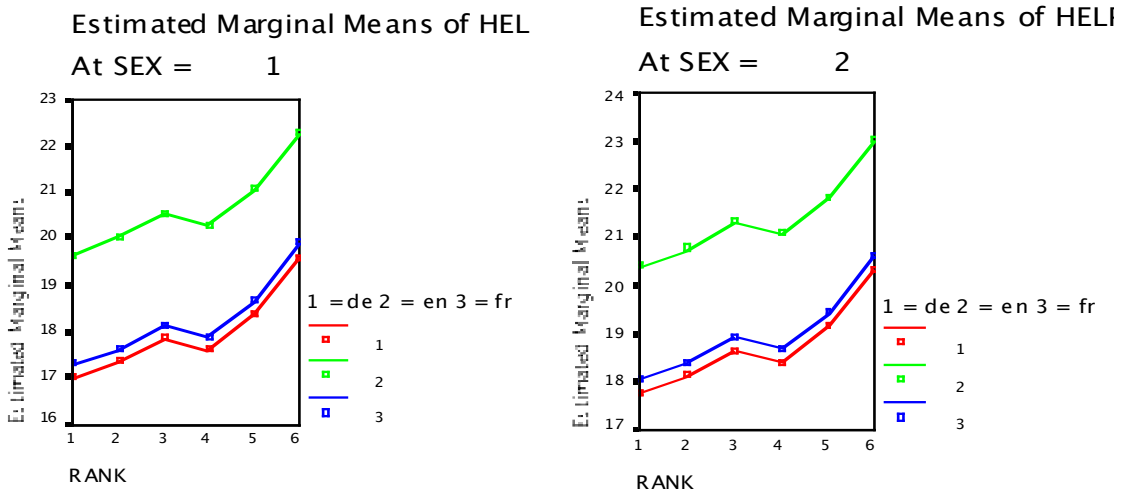
Remember that only SWB Sample 2 contains data from all three language-specific versions of the HBI, whereas SWB Sample 1 was all German, as were the samples collected by Frühauf, Steinhoff, and Hagge, of course. Thus, SWB2 provides us with the first chance to take a peek at national or cultural differences in terms of burnout. Unfortunately, but inevitably, those are inextricably confounded with linguistic differences.

The most conspicuous finding is that EN language means are almost everywhere highest for HBI scales and the CRISIS item (the exception being scale VOID, where DEs score highest). In contrast, DE and FR means are generally pretty close, with DE means being the second highest for six scales while FR means are the second highest for three scales.

This finding, although not reflecting a strong effect, looks pretty consistent at first glance. It holds for both genders and all six status categories and persists after controlling for age.

Fig. 2 provides a typical example, the dependent variable being scale HELPLESSNESS. It is also typical for the effect of status, to be discussed below.

Figure 2
HELPLESSNESS as a Function of Language, Status, and Sex



Legend: Left panel men, right panel women. Green curve = English, blue curve = French, red curve = German. Rank = status group (cf. Tab. 2)

Does the somewhat special position of the EN subgroup exist also at the item level? For 30 out of 40 HBI items, it does, again controlling for age, sex and status.

The effects of language on HBI scales and the CRISIS item are highly significant, but numerically very weak. Effect size coefficients *eta-squared* range from a low of .006 (for scales VOID and the CRISIS item) to .035 (for scale PERSONAL ACCOMPLISH-MENT). What is more impressive is the consistency across scales and the lack of interactions with sex and status.

Possible explanations, not mutually exclusive, include: (a) There is a genuine effect, i.e., users from English-speaking parts of the world tend to score higher on burnout; (b) the majority of EN respondents were from North America (probably mostly the US) and England, where living and working conditions may be more conducive to burnout than elsewhere[2]; (c) the translation into English is non-equivalent to the DE and FR version.

[2]I am indebted to Beate Schulze, Zurich, for this hunch.

In order to weigh the merits of explanation (b), separate ANCOVAs (with age as a covariate) were run, excluding all respondents who had indicated they were from North America (N = 373) or England (N = 108). Although that left less than half of the EN subsample, lines in the graphs tended to run parallel as before (with the exception of the CRISIS item, where complex interactions took place). And, although the differences tended to diminish somewhat, the EN curves again ran highest for six of the ten HBI scales (the remainder highlighting the DE subsample).

So, although there seems to be something to the North America/England hypothesis, explanations (a) and (c) cannot be ruled out.

4.4 Gender Differences

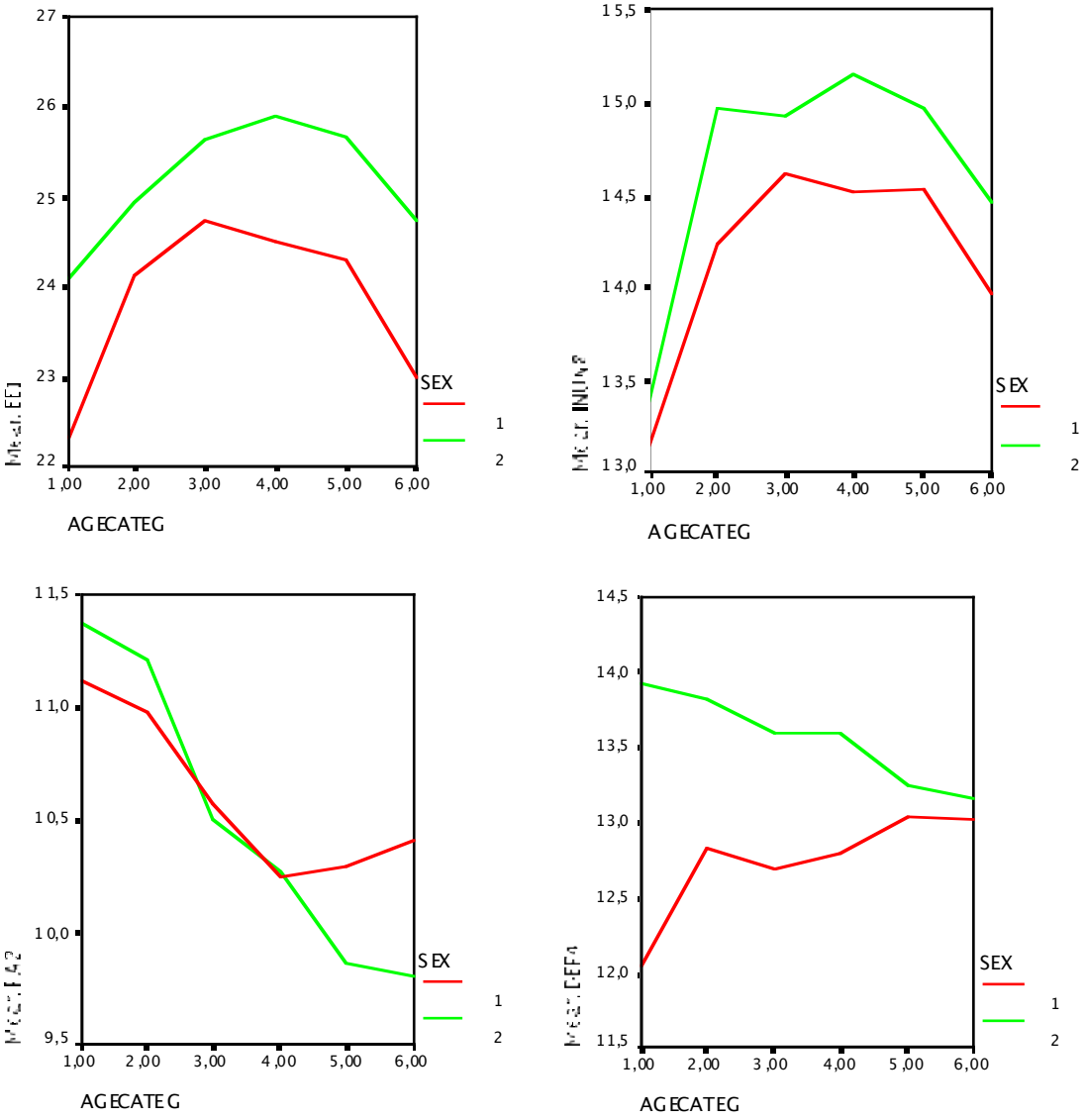
In accord with the gist of previous research (Rösing, 2003, p. 94-95), there was not much of a difference in the way men and women described themselves in terms of burnout. Eta^2 ranged from zero (for scale VOID and the CRISIS item) to .015 (for scale DISTANCING). Females scored somewhat higher on seven of the ten HBI scales; males scored somewhat higher on the remainder (namely, scales PERSONAL ACCOMPLISHMENT, DISTANCING, and TEDIUM). This held for all three languages and for all six status groups (age controlled). For the CRISIS item, there were interactions; thus, no clear picture emerged.

All of this, of course, may still be a selection effect. Use of the Internet at least up to 2007 differed for the sexes, and it probably differed differentially in various nations and language groups. Thus, the above differences, very small to begin with, may be explained that way.

4.5 Age Differences

Product-moment correlations between HBI measures and age were minimal, the maximum being -.08 for PA. However, grouping age into decades (<20 = 1; <30 = 2; ..... <70 = 6, omitting the 22 respondents over 69) and plotting means against age categories showed some interesting nonlinear trends, of two types. For scales EMOTIONAL EXHAUSTION, DISTANCING, HELPLESSNESS, VOID, INABILITY TO UNWIND, OVERTAXING ONESELF, AGGRESSION, and the CRISIS item, the curves were inversely U-shaped, with maxima in the twenties, thirties, or forties. See Fig. 3 for two examples (top panels). The other type, a more or less monotonic decrease with age, was found for PERSONAL ACCOMPLISHMENT and TEDIUM (see Fig. 3, lower left panel). For Scale DEPRESSIVE REACTION, the female curve was monotonically decreasing, the male curve more or less monotonically increasing (see Fig. 3, lower right panel). Note that the distances between maxima and minima are small, though.

Figure 3
Age Trend Examples for Four HBI Scales



Legend: Top left panel: EMOTIONAL EXHAUSTION; top right panel: INABILITY TO UNWIND; bottom left panel: PERSONAL ACCOMPLISHMENT; bottom right panel: DEPRESSIVE REACTION. Green curve: Female; red curve: Male. Abscissa: Age decades from teens to sixties.

4.6 Differences Between Occupational Status Groups

Although the same reservations apply here — executives who take the time to fill in the HBI on SWB's website may differ from their unemployed counterparts in more than one aspect — the sheer similarity of the plots for status (separately for language and sex, and

controlled for age) is stunning (for an example, cf. Fig. 2). With two exceptions, namely scales INABILITY TO UNWIND and OVERTAXING ONESELF, groups 6 („unemployed“) or/and 5 („other“) scored highest on the ten HBI scales and the CRISIS item. Of the nine cases in question, eight saw the Unemployed at the top.

How about minima? Well, that picture looks even clearer: Status group 1 („employee with executive functions“) scores lowest on eight of the ten HBI scales and the CRISIS item.

The exceptions: Scale INABILITY TO UNWIND sees status groups „Employee with Line Responsibility“ and „Other“ at the top, and „Employee without Line Responsibility“ at the bottom. With scale OVERTAXING ONESELF, we see „Independents“ up front, while „Employees without Line Responsibility“ seem to suffer least.

Although admittedly after the fact, that makes some sense at first blush.
Some attention to the „Other“ group may be warranted. After all, they scored second highest on nine out of eleven burnout indicators and highest on one (EE). Who are those Others?

I had expected to find students and homemakers, but that was definitely not the whole story. It turned out that the majority of the 1883 respondents who had assigned themselves to the *Other* category would have better fitted one of the defined categories. There were public servants („Beamte“), teachers, apprentices, consultants, firefighters and any number of jobs which in all likelihood were either employees without line responsibility or independent. But how to explain the conspicuously high burnout means?

After inspecting the job codes respondents had typed in (which included „blabla“ and „xxx“, of course), four clusters seemed sufficiently frequent to use them in an analysis: *students* (including pupils, doctoral students, and apprentices; N = 540), *housewives* (all who mentioned they were housewives or homemakers or mothers; N = 104), *teachers* (from Kindergarten teachers to school principals; all „education“; N = 343), and *public servants* („Beamte“; including many police officers; N = 93).

Thus, those four subcategories of „Other“ were compared with the rest of the „Others“ and with the vast remainder of SWB Sample 2 which includes categories 1-4 and 6 (cf. the list in Tab. 3).

The results supported two hunchs this writer has long held: Two categories of burnout victims go largely unnoticed, namely housewives and students (in the wide sense). Housewives (by definition all female; there was only a handful of housemen) scored

highest in their gender category on scales EMOTIONAL EXHAUSTION, DEPRESSIVE REACTION, HELPLESSNESS, VOID, TEDIUM, AGGRESSIVE REACTION, and the CRISIS item, often markedly. Students (both male and female) scored highest on scales PERSONAL ACCOMPLISHMENT (reversed) and OVERTAXING ONESELF. Moreover, male students obtained the highest means among other males for scales DEPRESSION, HELPLESSNESS, TEDIUM, and the CRISIS item.

Also ran and got to the top of their gender field: male teachers for scale EMOTIONAL EXHAUSTION, and teachers of both genders for INABILITY TO UNWIND. Again, this does not seem too surprising.

Of course, we cannot rule out the possibility that it was just the most desparate housewives who took the HBI, whereas all the others live happily and will do so ever after. To study questions as these, other — much more costly — research approaches are indicated.

5. Conclusions

The HBI appears to be a promising instrument for assessing individual and group levels of burnout. In view of the very short scales, reliabilities are adequate and validities mostly adequate.

This study explored for the most part one large international online sample. What emerged, however, cannot provide more than hunches, given the non-random character of the sample and its sub-samples.

6. References

Aronson, E., Pines, A. M., & Kafry, D. (1983). *Ausgebrannt.* Stuttgart: Klett-Cotta.

Borkenau, P. & Ostendorf, F. (1993). *NEO-Fünf-Faktoren Inventar (NEO-FFI) nach Costa und McCrae.* Göttingen: Hogrefe.

Burisch, M. (1984a). Approaches to personality test construction: A comparison of merits. *American Psychologist, 39,* 214-227.

Burisch, M. (1984b). *The Maslach Burnout Inventory and the Tedium Measure: Reliability and validity in a German sample.* Unpubl. ms.

Burisch, M. (1997). Test length and validity revisited. *European Journal of Personality,* 303-315.

Demerouti, E. (1999). *Burnout. Eine Folge konkreter Arbeitsbedingungen bei Dienstleistungs- und Produktionstätigkeiten.* Berlin: Peter Lang.

Ebbinghaus, M. (1986). *Erfassung von Burnout. Entwicklung und Überprüfung eines Meßinstrumentes für die Anwendung in verschiedenen Berufsbereichen.* Unpubl. thesis, U. Oldenburg, Germany.

Enzmann, D. (1996). *Gestresst, erschöpft oder ausgebrannt?* Munich: Profil.

Fahrenberg, J. & Selg, H. (1973). *Das Freiburger Persönlichkeitsinventar.* Göttingen: Hogrefe (2nd ed.).

Fahrenberg, J., Selg, H. & Hampel, R. (1984). *Das Freiburger Persönlichkeitsinventar FPI. Revidierte Form FPI-R und teilweise geänderte Fassung FPI-A1.* Göttingen: Hogrefe.

Frick, U. & Filipp, G. (1997). *Gesundheitsberufe im Land Salzburg. Berufs- und Lebenssituation.* Salzburg, Austria: Amt der Salzburger Landesregierung.

Frühauf, F. (1990). *Entwicklung eines deutschen Burnout-Inventars.* Unpublished thesis, University of Hamburg.

Hagge, M. (2005). *Vergleich zweier deutscher Burnout-Inventare.* Unpublished thesis, University of Hamburg.

Hotter, E. (2009). Personal communication.

Hotter, E. (2014). Personal communication.

Hotter, E. (2016). Personal communication.

Maslach, C. & Jackson, S.E. (1986). *Maslach Burnout Inventory Manual (2d. ed.)*. Palo Alto, CA: Consulting Psychologists Press.

Maslach, C., Jackson, S.E. & Leiter, M.P. (1996). *Maslach Burnout Inventory Manual (3d. ed.)*. Palo Alto, CA: Consulting Psychologists Press.

Mielke, M. (2014). *Die Validität des Hamburger Burnout-Inventars im Klinischen Kontext*. Unpublished thesis, University of Hamburg.

Rösing, I. (2003). *Ist die Burnout-Forschung ausgebrannt?* Heidelberg: Asanger.

Schaarschmidt, U. & Fischer, A.W. (2008). *AVEM. Arbeitsbezogenes Verhaltens- und Erlebensmuster*. London: Pearson.

Schaufeli, W.B. & Enzmann, D. (1998). *The burnout companion to study & practice.* Chichester: Taylor & Francis.

Sijtsma. K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107-120.

Tolke, A.-M. (2013). *Erweiterung eines bestehenden Burnout-Inventars. Effekte auf Reliabilität und Validität.* Unpubl. Thesis, U. of Kiel.

von Herder, F. (2011). *Das Hamburger Burnout-Inventar (HBI). Eine Studie zu Retestreliabilität und Korrelaten.* Unpublished thesis, University of Hamburg.

Weber, U. (2014). *Burnout-Prävention im Internet. Konzeption und Evaluation eines Online-Präventionsprogramms.* Unpubl. Dissertation, U. Hamburg.

Winer, B. J. (1962). *Statistical principles in experimental design.* New York: McGraw-Hill.

Wurm, W., Vogel, K., Holl, A., Ebner, C., Bayer, D., Mörkl, S., et al. (2016). Depression-Burnout overlap in physicians. *PLoS ONE 11(3):* e0149913. doi:10.1371/journal.pone.0149913