Long-term trends (1980–2017) in the N-pact factor of journals in personality psychology and individual differences research

Michael Kossmeier, Johannes Vilsmeier, Rosalie Dittrich, Tanja Fritz, Caroline Kolmanz, Constantin Y. Plessen, Agnieszka Slowik, Ulrich S. Tran, Martin Voracek

Angaben zur Veröffentlichung / Publication details:

Kossmeier, Michael, Johannes Vilsmeier, Rosalie Dittrich, Tanja Fritz, Caroline Kolmanz, Constantin Y. Plessen, Agnieszka Slowik, Ulrich S. Tran, and Martin Voracek. 2019. "Long-term trends (1980–2017) in the N-pact factor of journals in personality psychology and individual differences research." *Zeitschrift für Psychologie* 227 (4): 293–302. https://doi.org/10.1027/2151-2604/a000384.

Original Article

Long-Term Trends (1980–2017) in the *N*-Pact Factor of Journals in Personality Psychology and Individual Differences Research

Michael Kossmeier, Johannes Vilsmeier, Rosalie Dittrich, Tanja Fritz, Caroline Kolmanz, Constantin Y. Plessen, Agnieszka Slowik, Ulrich S. Tran, and Martin Voracek

Department of Basic Psychological Research and Research Methods, School of Psychology, University of Vienna, Austria

Abstract: Recent metascience investigations of the *N*-pact factor (NF; median sample size of studies published in a journal) have revealed NFs of merely about 100 in fields like social, sport, and exercise psychology. Journal NF has also been shown to correlate negatively with journal impact factors (JIF), implying that smaller studies appear in more prestigious journals. In this first long-term and largest NF analysis to date (3,699 articles coded), annual NFs of two personality psychology journals were tracked over 38 years since their inception in 1980. Overall NF was about 190, gradually increased over time, and within-journal NF-JIF correlations were positive. Online samples and articles featuring supplemental files presented larger NFs, whereas those involving student samples had smaller ones. Sample size distributions showed multimodality, and surplusses of even-numbered sample sizes and of those just beyond 100 were evident. An NF statement, accompanying authors' submitted papers, is suggested.

Keywords: effect size, *N*-pact factor (NF), power, sample size, time trends

The current decade has witnessed widespread skepticism (aka the replicability debate) about the trustworthiness and reproducibility of published empirical research findings across many fields, including psychological science (Open Science Collaboration, 2015; Simmons, Nelson, & Simonsohn, 2011). This fundamental issue has led to large-scale replication initiatives, changes in journal policies, emergence of an Open Science culture, reform movement in research methods and statistics, and development of statistically based diagnostic tools for assessing the evidentiality (i.e., evidential value) of research findings. Psychological science clearly is at the forefront of these important developments and advancements (Nelson, Simmons, & Simonsohn, 2018).

Among the latter innovations (of evidentiality-testing procedures), the *N*-pact factor (NF) has been defined as the median value of the distribution of sample sizes of empirical studies published in one year in one journal (Fraley & Vazire, 2014). Using prior knowledge of field-typical effect sizes, the NF allows estimating the typical analytic power of empirical studies for a specified frame of interest (journals, researchers, research designs, topics, institutions, countries, subdisciplines, etc.). For this reason, the NF is an important quality criterion of journals and a valuable alternative to the much-criticized journal impact factor (JIF).

Initial metascientific accounts, introducing the NF, have yielded evidence for worryingly small sample sizes (and, consequently, underpowered studies) across leading journals in social and/or personality psychology (Fraley & Vazire, 2014). Specifically, an NF analysis of empirical papers published in six such journals (*Journal of Personality*, *Journal of Research in Personality*, *Personality and Social Psychology Bulletin*, *Journal of Personality and Social Psychology*, *Journal of Experimental Social Psychology*, along with *Psychological Science*) during the period 2006–2010 overall found an NF of merely slightly above 100. Added to this, worrisome patterns of negative correlations between NF and JIF emerged: that is, the supposedly "best" journals (according to their JIF) on average published the smallest, most underpowered, studies. Relatedly, an NF analysis of empirical papers published in four leading journals in sport and exercise psychology (*International Journal of Sport Psychology*, *Journal of Applied Sport Psychology*, *Journal of Sport and Exercise Psychology*, *Psychology of Sport and Exercise*) during the period 2009–2013 yielded an only slightly higher average NF of 114 (Schweizer & Furley, 2016). Most recently, an NF analysis of two leading APA (American Psychological Association) journals in clinical psychology (*Journal of Abnormal Psychology*, *Journal of Consulting and*

Clinical Psychology), based on four separated publication years (2000, 2005, 2010, and 2015), yielded a noticeably higher overall NF of 179 (Reardon, Smack, Herzhoff, & Tackett, 2019).

As important as these three pioneering accounts on the NF are, a number of limitations and disparities, as well as remaining research gaps and unknowns, are obvious. First, the social/personality and clinical psychology NF studies (Fraley & Vazire, 2014; Reardon et al., 2019) did not distinguish between study designs, which constitutes an important omission and source of obfuscation. Although the sport/exercise psychology NF study (Schweizer & Furley, 2016) did distinguish between correlational, quasi-experimental, and experimental designs, more fine-grained distinctions (between-subject vs. within-subject vs. mixed design types) were only made within the experimental design category. Second, the sport/exercise psychology NF study did not investigate NF-JIF correlations.

Third, and most important of all, although two of the preceding NF studies scrutinized about a handful of journals (or two; Reardon et al., 2019), all three existing NF analyses either had investigation periods that were rather short (covering merely 4–5 publication years each) or temporally separated (Reardon et al., 2019) and, moreover, were quite recent (social/personality psychology: 2006–2010; sport/exercise psychology: 2009–2013; clinical psychology: 2000, 2005, 2010, 2015). Indeed, these recent investigation periods coincide with the onset of the awareness of a replication and confidence crisis in empirical science (2011 at latest) and partly even extend into those years, where already formative actions had been taken against the perceived crisis (early to mid-2010s). Specifically, the publication year 2008 of two journals (*Journal of Personality and Social Psychology*, *Psychological Science*) covered in the social/personality psychology NF study overlaps with the study sampling frame of the Reproducibility Project: Psychology (Open Science Collaboration, 2015). Hence, it is obvious that extant NF studies in some measure either describe articles (partly possibly questionable) which gave rise to the current concerns about the trustworthiness of published empirical research, or articles submitted and accepted during the most recent years, when there already was awareness of, and counter-measures against, the replicability and confidence crisis.

However, at the same time all three existing NF studies do not provide a deeper historical perspective: little is known about possible long-term trends in the NF. In this context, it is important to recall that, from the earliest investigations of the statistical power of published journal articles in psychological science onwards (Cohen, 1962), it has been stressed that studies invariably are underpowered (with estimates in the range of 35–50%), with no ameliorating time trends discernible. If so, this would be a curious fact, thinking of the enormous facilitations relating to the planning, collection, management, analysis, reporting, and provision of empirical research data that happened over the past few decades.

In a nutshell, prior to the 1980s, data management and analysis was a matter of mainframe computers, punchcards, and computer programmers. Personal computers came up in the early 1980s, the influential statistical software package SPSS is available with graphical user interface (Windows) since the early 1990s, collecting data online is feasible since the late 1990s, and supplemental files in journals gained momentum in the 2000s. It is important to note that tabulations of statistical power for appropriate sample size planning have been accessible since much earlier on (Cohen, 1969, 1988), and user-friendly power-analysis software is available since the mid-1990s (Erdfelder, Faul, & Buchner, 1996).

To investigate for the first time long-term trends in the NF, we made use of the serendipitous opportunity of two equally old, long-standing journals from the same psychological subfield (personality and differential psychology), both of which show many similarities, but also some differences (detailed below). This allowed us to address the above-mentioned limitations and knowledge gaps of extant NF studies.

Methods

Sampling Frame

Both the *Journal of Individual Differences* and *Personality and Individual Differences* (henceforth, JID and PAID) were founded in 1980. JID, under its former name *Zeitschrift für Differentielle und Diagnostische Psychologie* (which translates to: *Journal of Differential Psychology and Psychological Assessment*), was a German-language journal before it was internationalized (Hennig, 2005). Its publishing volume has more or less been constant over time (about 30 articles/year), whereas PAID, in recent years, has increased its publishing volume dramatically (i.e., many times over).

We analyzed the entirety of JID articles (about 1,000 during the 38 publication years 1980–2017), along with an annually stratified proportional random sample (25% of articles/year, but 50 articles/year as a minimum; totaling well over 3,000 articles) of the about 10,000 PAID articles published during the same period.

Coding Form for Study Variables

Drawing on procedural details of extant NF studies and extending these, we recursively developed a coding form

(along with annotations and definitions for coders), which we tested and refined with a small training sample of articles which was processed by six independent coders (authors AS, CK, CYP, JV, RD, and TF). These trained (precalibrated) coders were then randomly allocated to the corpus of articles. For non-eligible (non-empirical, meta-analytic, and other) articles, we recorded the exclusion reason in detail, differentiating more than a dozen types of non-eligible published material. The units of analysis were study samples; consequently, multi-study or multi-sample papers contributed more than one datapoint for the NF analyses.

Following Schweizer and Furley (2016), we differentiated correlational from quasi-experimental and experimental study designs and, within the latter two categories, between-subject from within-subject and mixed designs. In addition, we recorded whether samples were collected online, whether they comprised university students, and whether articles had supplemental files.

Disclosure of Open Science Practices

"We disclose how we determined our sample size, all data exclusions, all manipulations, and all measures in the study" (Simmons, Nelson, & Simonsohn, 2012, p. 4). For this metascience study (i.e., research about research), sample size was arrived at through the predefined sampling frame (as detailed above). Data exclusions followed from the above eligibility criteria. There were no experimental manipulations. All statistical manipulations are indicated alongside the respective analyses. Apart from the measures listed in the coding form, which comprised the dataset underlying all analyses, no further measures were collected.

We did not preregister this metascientific investigation; thus all reported analyses and findings should be considered as exploratory and descriptive. All components required for reproducible data analysis are archived at the Open Science Framework (annotated coding form: https://osf.io/kcuyj/; dataset for analysis: https://osf.io/nrmh3/; code to reproduce the analyses: https://osf.io/yvnp6/), a research repository compliant with the FAIR (findable, accessible, interoperable, re-usable) guiding principles for scientific data (Wilkinson et al., 2016).

Results

Sampling Frame and Sample Descriptive Statistics

All 976 articles published 1980–2017 in JID and an annually stratified proportional random sample (specified as detailed above) of 2,723 articles from the total of 9,910 articles published 1980–2017 in PAID were included in analysis, yielding a total of 3,699 articles considered for NF analysis. Of these, 526 articles (14.2%) were not informative for NF analysis and therefore excluded (see Table 1 for a breakdown of exclusion reasons; e.g., the current account would not go into such an NF analysis, because it is a metascience study). Of the remaining 3,173 articles, 19.8% (139 out of 702) from JID and 14.3% (354 out of 2,471) from PAID contained more than one sample, providing 912 (JID) and 2,999 (PAID) samples, and a total of 3,911 samples, available for NF analysis.

Table 1. Distribution of reasons for exclusion of articles not eligible for the NF analysis

Reason for exclusion	Frequency (%)
Theory/conceptual paper	94 (17.9%)
Methodology paper	89 (16.9%)
Review	71 (13.5%)
Editorial	68 (12.9%)
Letter/commentary/reply	60 (11.4%)
N equivocal	51 (9.7%)
Book review	50 (9.5%)
Meta-analysis	17 (3.2%)
Published erratum/correction	12 (2.3%)
Qualitative study	7 (1.3%)
Metascience paper	3 (0.6%)
Single-case study/case series	2 (0.4%)
Systematic review	2 (0.4%)

Note. Frequencies and percentages of exclusion reasons for all excluded articles (N = 526), calculated for JID and PAID combined. NF = *N*-Pact Factor; JID = *Journal of Individual Differences*; PAID = *Personality and Individual Differences*.

Time Trends in the NF and Journal Comparisons

Overall NFs amounted to 185 (JID) and 192 (PAID), that is, when calculated across the entire observation period of 38 publication years (1980–2017). Decade-wise partitioning of NF values (calculated for 1980–89, 1990–99, 2000–09, and 2010–17) suggested a gradual increase over time: NF values were 155, 143, 228, and 194 for JID; and 106, 132, 200, and 262 for PAID. Annual NF values were positively associated with publication year (Spearman r_s = .40 and .83 for JID and PAID, respectively; see Figure 1 for the smoothed time-series data).

Across the entire timespan, PAID had somewhat higher NF values than JID (median difference of annual NF values: 12.5). Annual NF differences between JID and PAID were negatively associated with publication year (r_s = −.43), suggesting that the NF of PAID grew at a faster rate than the NF of JID.

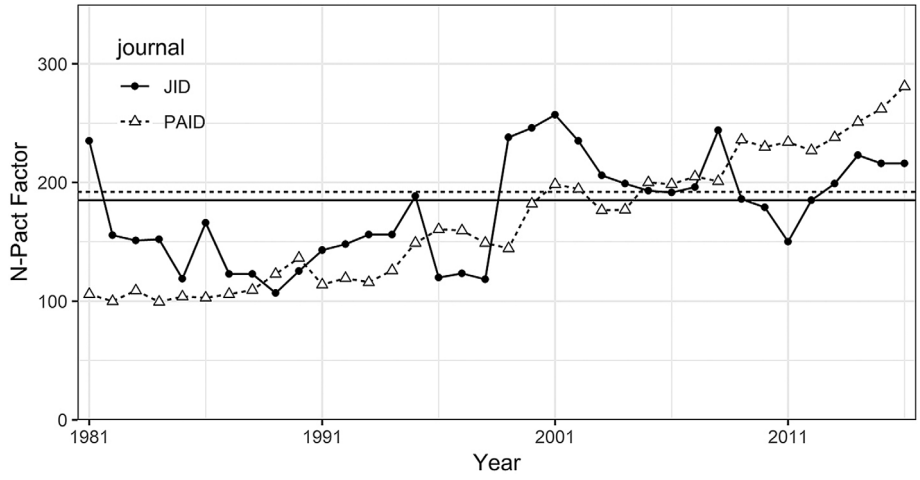

Figure 1. Time trends in the NF of journals in personality and individual differences research (1980–2017). Annual median sample size (the journal *N*-pact factor [NF]) for articles published in the *Journal of Individual Differences* (JID) and *Personality and Individual Differences* (PAID). For visualizing the upward trend more clearly, the time-series data were smoothed through applying 3-year moving medians (i.e., the value for a specific year was calculated from that year, along with the immediately preceding and the immediately following year). Horizontal reference lines are set at the grand total NF (185 for JID, 192 for PAID) over the entire timespan. *R* code to reproduce the figure is available at https://osf.io/rbdkz/

Associations Between NF and Journal Impact Factor (JIF)

For JID and PAID, 2-year JIFs were available from 2010 and 1997 onwards, respectively; and 5-year JIFs from 2013 and 2007 onward, respectively. For both journals, there was a correspondence between NF and 2-year JIF (JID: r_s = .71; PAID: r_s = .75), as well as between NF and 5-year JIF (JID: r_s = .80; PAID: r_s = .63). Concomitantly, the journals' relative standing within the respective Web of Science journal category ("Psychology, Social") improved: year and within-category JIF percentile were associated with r_s = .41 (JID) and .70 (PAID).

Prevalence of Study Designs and NF-Derived Power Estimates

As expected for the subfield of personality psychology and individual differences research, correlational studies by a large margin were the most-commonly occurring study format appearing in both JID and PAID articles (71.9% and 69.3%, respectively), followed by quasi-experiments (21.4% and 25.3%), whereas true experimental designs were rare (6.7% and 5.2%). These proportions were similar for JID and PAID and moreover rather stable over time (Figure 2A). Study-design specific NF values were highest for correlational designs, whereas considerably lower for quasi-experimental and experimental designs (Table 2).

Study-design specific power analysis (Figure 3) applied to the NF values (following the specifications of Schweizer & Furley, 2016) showed that correlational designs would have sufficient power (80%) to detect small-to-medium effects (r = .20), but would be underpowered for the detection of small effects (r = .10). Quasi-experimental designs would be sufficiently powered to detect medium-sized effects (r = .30), whereas experimental designs proper had the lowest NF and thus the lowest analytic power.

NF Associations With Online and Undergraduate Samples, Articles With Supplemental Files, and Journal Internationalization

Sample size distributions were noticeably positively skewed (i.e., presenting more smaller than larger samples). For the following group comparisons, we thus used the Mann-Whitney U test, and calculated from its z statistic the nonparametric effect size D (Kraemer & Andrews, 1982), which is analogous to Cohen d, but unaffected by distributional skewness and outliers. As an additional effect-size indicator for the magnitude of these group differences, we calculated the probability of superiority measure PS, representing the probability that a randomly sampled datapoint from one group has a higher score than a randomly sampled datapoint from the other group (Grissom, 1994).

The first online samples appeared in 1998 (JID) and 1996 (PAID). Online samples were larger than conventionally collected samples published from 1998 onward in JID (NF = 359 vs. 191; Mann-Whitney U test: z = 4.62, D = 0.64, PS = .68), or published from 1996 onward in PAID (NF = 296 vs. 209; z = 5.24, D = 0.31, PS = .59).

The first supplemental files appeared in 2015 (JID) and 1990 (PAID). Samples in articles featuring additional information via supplemental files were larger than those without published from 1990 onward in PAID (NF = 262 vs.

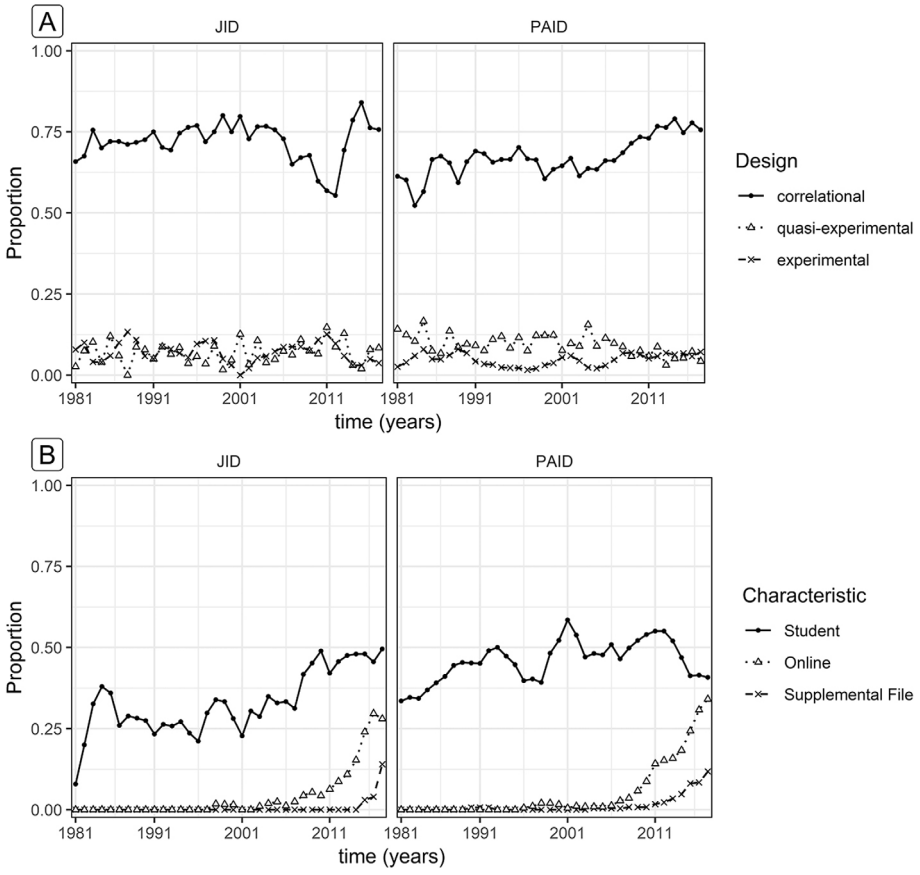

Figure 2. Time trends in study characteristics (1980–2017). The line charts show annual proportions of selected study characteristics over time and differentiated by journal. For visualizing the time trends more clearly, the time series data were smoothed through applying 3-year moving medians (i.e., the value for a specific year was calculated from that year, along with the immediately preceding and the immediately following year). (A) Proportion of study designs (correlational, quasi-experimental, and experimental). (B) Proportion of online samples, student samples, and articles with supplemental files. *R* code to reproduce the figure is available at https://osf.io/ayce2/

204; $z = 2.88$, $D = 0.29$, $PS = .59$), but not for articles published from 2015 onward in JID (NF = 212.5 vs. 223; $z = -0.22$, $D = -0.06$, $PS = .52$).

University student samples comprised a noteworthy share of the empirical articles included in the NF analysis (JID: 35.5%; PAID: 45.5%). Perhaps surprisingly, such easily collected convenience samples typically were smaller, instead of larger, when compared with the remainder (non-undergraduate, community, and general population) of samples. This was true for JID (NF = 121 vs. 256; $z = -9.22$, $D = -0.64$, $PS = .68$), as well as for PAID (NF = 159 vs. 231; $z = -9.91$, $D = -0.36$, $PS = .60$). Student samples were only trivially less likely collected online, as compared to non-student samples (JID: $r = -.03$; PAID: $r = -.05$). Also, the NF of student samples was consistently smaller than the NF of non-student samples, when

Table 2. NF values differentiated by study design

Study design	Number of samples (%)		NF (1980–2017)	
	JID	PAID	JID	PAID
Correlational	656 (71.9)	2,082 (69.4)	220	225
Quasi-experimental (between)	163 (17.9)	652 (21.7)	124	134
Quasi-experimental (within)	6 (0.7)	24 (0.8)	79	104
Quasi-experimental (mixed)	26 (2.9)	84 (2.8)	131	82
Experimental (between)	30 (3.3)	110 (3.7)	78	93.5
Experimental (within)	12 (1.3)	25 (0.8)	81	102
Experimental (mixed)	19 (2.1)	22 (0.7)	67	101

Note. Frequencies and percentages of different study designs, as appearing in JID and PAID across the entire observation period, and study-design specific NF values. JID = *Journal of Individual Differences*; PAID = *Personality and Individual Differences*.

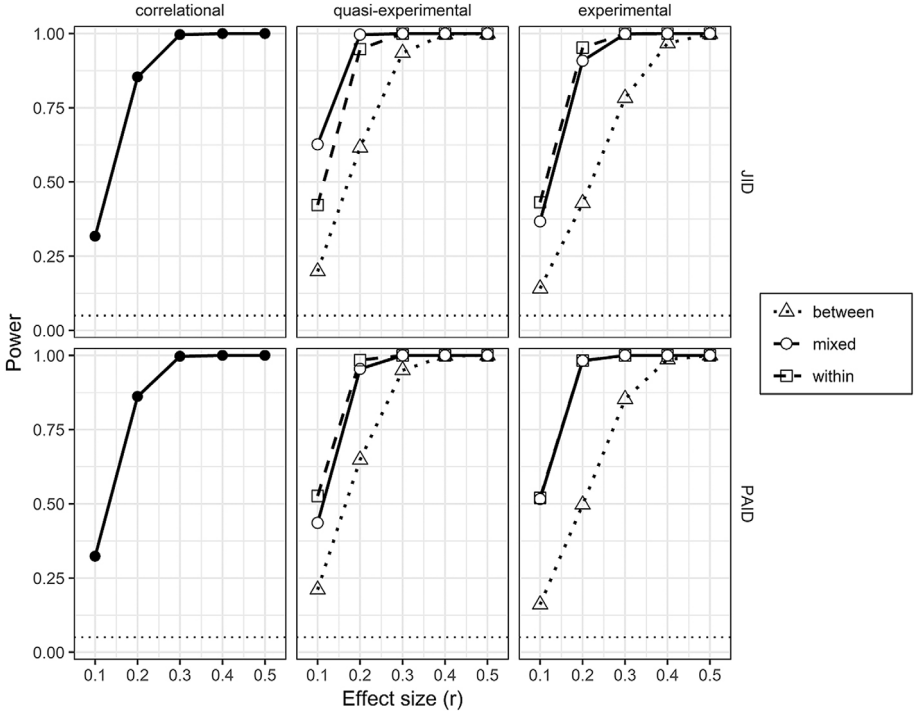

Figure 3. Study-design specific power values corresponding to the NF values. Shown are the values of analytic power, corresponding to the observed NF values and for $\alpha = .05$ (two-tailed), for different scenarios of the magnitude of true effects ($r = .10, .20, .30, .40$, and $.50$). Power values were calculated (see Schweizer & Furley, 2016) for a t test of whether the true correlation is unequal zero (for correlational designs), an independent-groups t test (for between-subject designs), a dependent-groups t test (for within-subject designs), and an F test of the interaction effect in a 2 × 2 mixed-model analysis of variance (ANOVA) design (for mixed designs). *R* code to reproduce the figure is available at https://osf.io/2ktcj/

analyzed separately for samples collected online (JID: NF = 243 vs. 399; $z = -2.83$, $D = 0.82$, $PS = .74$; PAID: NF = 284.5 vs. 308.5; $z = -1.52$, $D = 0.17$, $PS = .55$) versus samples not collected online (JID: NF = 120 vs. 243; $z = -8.66$, $D = 0.62$, $PS = .68$; PAID: NF = 148 vs. 222; $z = -9.46$, $D = 0.37$, $PS = .61$). Therefore, a smaller proportion of online samples among student samples cannot serve as a viable explanation for the unexpected size differences observed between student and non-student samples.

For both journals, the proportions of online samples and supplemental files were on the rise during recent years (Figure 2B). In addition, the proportion of student samples also increased over time for JID: the occurrence of student samples was positively associated with publication year ($r_s = .19$), but not for PAID ($r_s = .02$). Finally, the NF of JID was somewhat higher after its internationalization in 2005 (1980–2004 NF: 168, vs. 2005–17 NF: 196; $z = 2.89$, $D = -0.19$, $PS = .56$).

Distributional Characteristics of Sample Sizes

The highly positively skewed sample size distribution (for JID and PAID combined, and truncated for either $N >$ 500 or $N > 110$) is displayed in Figure 4A and B. Visual inspection and further analysis revealed that the sample size distribution showed multimodality (most prominently at $N = 100$, 40, and 60; all divisible by 10), and that sample size numbers in general showed certain digit preferences and a clustering around "nice, round numbers."

Even-numbered sample sizes outnumbered odd-numbered ones (2,228 out of 3,911 samples, or 57%; binomial test: $z = 8.71$). The sample size distribution also showed a peculiarity in the vicinity of $N = 100$: we applied the caliper test (Masicampo & Lalande, 2012), originally proposed to test for a surplus of p values just below the conventional significance threshold ($p < .05$), to test for a disparity between the 10 smallest three-digit sample sizes (i.e., $N = 100$–109) versus the 10 largest two-digit sample sizes (i.e., $N = 90$–99). Sample sizes just beyond $N = 100$ were more prevalent than sample sizes just below $N = 100$ (184 out of 316, or 58.2%; binomial test: $z = 2.93$).

Proportion of False Positives

Estimating the proportion of false-positive findings is feasible by assuming that: (i) all studies report statistically significant findings, (ii) questionable research practices

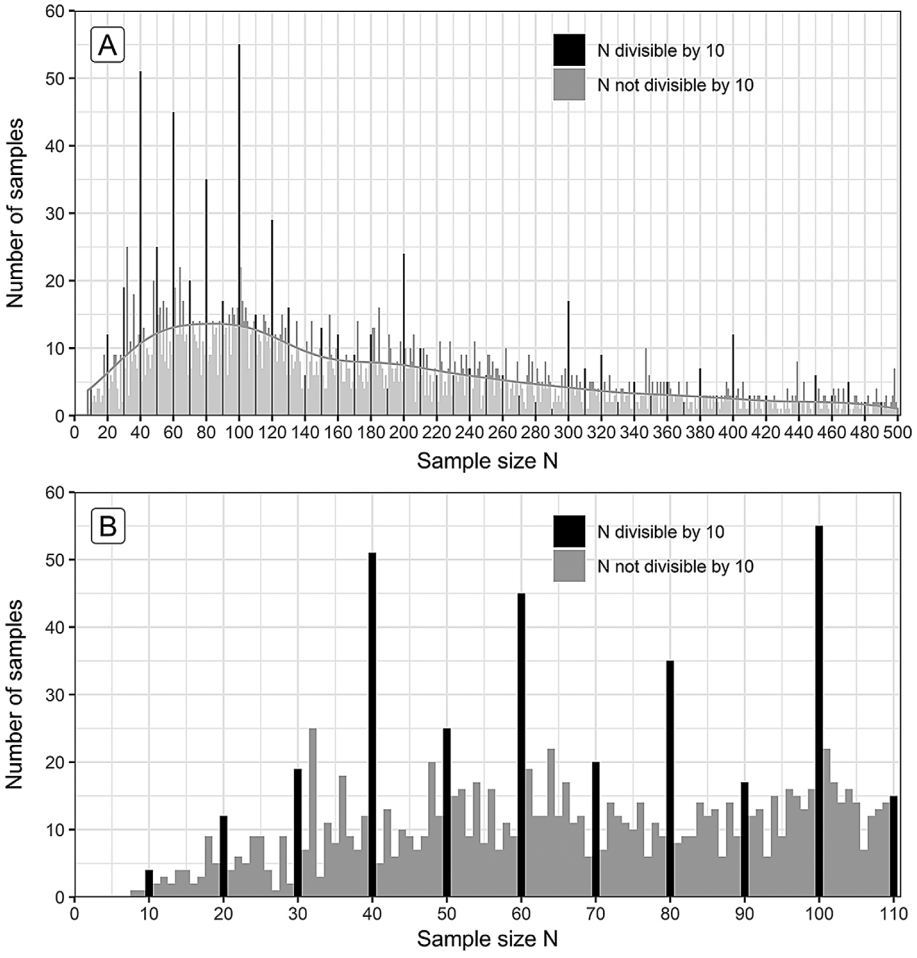

Figure 4. Distribution of reported sample sizes. Needle plots, displaying the number of samples with a specific sample size (JID and PAID combined), along with a scaled density estimate. Highlighted (black needles) are sample sizes divisible by 10. (A) Sample size distribution, truncated at $N > 500$. (B) Zoomed-in version of (A), truncated at $N > 110$, highlighting the distributional characteristics over the range of smaller ($N < 100$) sample sizes and in the vicinity of $N = 100$. *R* code to reproduce the figure is available at https://osf.io/u6xhb/

(*p*-hacking, and other) are absent, and (iii) the prevalence of true null hypotheses is known. Of these three assumptions, (i) for all practical purposes is true (Fanelli, 2012; Open Science Collaboration, 2015); (ii) in all likelihood is not true (John, Loewenstein, & Prelec, 2012); and although (iii) really is unknown, different scenarios can be tested.

For estimating the rate of false positives, we calculated the typical power (i.e., the median value) of JID and PAID studies from the power of individual studies, thereby accounting for the different study designs (as tabulated in Table 2). For small-to-medium true effects ($r = .20$) and $\alpha = .05$ (two-tailed), the resulting typical power of studies would be very similar for both journals: 83% (*IQR*: 54.1–98.9%) for JID studies and 82.7% (*IQR*: 52.4–99%) for PAID studies. Factoring in these typical power values in the false-positive calculations, and further assuming prevalences of true null hypotheses of 20%, 50%, and 80% (along with no *p*-hacking), led to false-positive estimates of 1.5%, 5.7%, and 19.4% (JID), and of 1.5%, 5.7%, and 19.5% (PAID).

Discussion

We provide a bird's eye view (i.e., a metascientific inquiry) on the trajectory of the NF for two (partly comparable, partly different) journals in the field of personality psychology and individual differences research. We backtrace time trends observed in the NF for these two journals for almost four decades, since their inception in 1980. Thereby, we add a necessary historical perspective to extant research related to current journal NFs in psychological science. Further, with 3,700 journal articles manually coded in detail, our investigation is the largest NF analysis to date.

This first analysis of long-term time trends in the NF, based on two journals covering the same subfield of psychological science, generated a wealth of novel, informative, and partly surprising insights into the development of study sample sizes over time. Early on in the current replicability debate, it had already been conjectured that personality psychology and individual differences research

might be less affected by underpowered studies than social psychology is (Asendorpf et al., 2013). This is indeed suggested by the current findings: NF figures were almost twice as high than those previously reported for sport and exercise psychology journals (Schweizer & Furley, 2016), or social psychology journals in particular (Fraley & Vazire, 2014), while being quite similar to the NF of prestigious journals in clinical psychology (Reardon et al., 2019).

This specific pattern of findings across existing NF studies opens up several fruitful avenues for future research along these lines: among others, further NF analyses could aim to identify further high-NF versus low-NF subfields within psychological science. A related question to address is whether different subfields, on average, investigate effects of similar or different magnitude: if, for example, effects commonly investigated in social psychology are indeed larger than elsewhere, social psychological research might not be underpowered; if these are not larger, it would be underpowered.

One key finding was (contrary to notions about stagnating statistical power of psychology studies) a positive time trend in the data: NF values gradually increased over time, as has been observed in other fields (Lamberink et al., 2018). While overall NF values were about 190 for both JID and PAID, more recent (2000s) and current (2010s) NF values amounted to 200, or even to 250. All of this emphasizes the necessity of further time-trend investigations into the NF. One specific direction for future research would be to compare the development of NFs across different subfields: for example, has the NF of social psychological research increased since the 1970s or 1980s to a similar degree than the NF of personality research, or less so, or not at all?

NF-JIF associations throughout were markedly positive, which was true for both journals. This evidence of correspondence means that, with JIFs increasing over time, sample sizes of studies published in these journals also became larger. Concomitantly, the journals' relative standing (as indicated by their JIF) in their field increased. This differs from evidence from social psychology journals in particular, where, ironically, the journals with the highest JIFs published the smallest studies and thus presented the lowest NFs (Fraley & Vazire, 2014). Although still heavily marketed, the JIF is one widely criticized, easily manipulable prestige marker for journals, not a quality criterion of individual articles (Brembs, Button, & Munafò, 2013), whereas the NF is one quite straightforward quality criterion of empirical research. Bearing this in mind, future research, covering other journal families and science fields (Szucs & Ioannidis, 2017), should test whether NF-JIF correlations are directionally as desired, that is, substantially positive, such that sufficiently powered large-*N* studies preferably are published in high-JIF outlets.

All in all, we observed more similarities than differences between the two journals under scrutiny. An important limitation is that it remains to be seen whether the general trends observed here generalize to the broader journal landscape of personality and differential psychology. Less than 15% of the total journal contents were not empirical primary studies and thus not amenable for NF analysis. The breakdown of main categories of study designs showed that, as expected for this subfield, correlational designs prevailed by a large margin (70% of studies), whereas comparisons of natural groups (quasi-experimental designs) had a share of about 25%, and actual experimental research was infrequent (about 5%). As well, this breakdown seemed to be stable over time.

NF values were highest for correlational designs, indicating that these studies would have sufficient power (about 80%) to detect even small-to-medium effects (r = .20). Paralleling prior related research on the NF (Schweizer & Furley, 2016), power estimations were lower for quasi-experimental designs, and lowest for experimental research proper. In this context, the standard argument is that experimental research might commonly investigate larger effects than non-experimental research and for this reason in truth might not be underpowered. However, we would like to caution against any hasty conclusions and rather suggest that precisely this claim itself, namely, that experimental research supposedly probes larger effects than non-experimental research, is in need of empirical substantiation.

An additional point of consideration is that the design-specific power estimates derived from the NF are only valid for standard cases, namely, the simplest implementation of the respective study designs (see Figure 3 note). To the extent that in correlational studies correlations are calculated within subgroups, that more than two groups are compared in quasi-experiments, or that an experiment is more sophisticated than a 2 × 2 design, power would be lower and thus overestimated, when derived from NF analysis (Schweizer & Furley, 2016).

Accounting for the different design types, the typical study power for both journals amounted to slightly above 80%, which is satisfying. Even assuming that only 50% of research hypotheses are true (stated differently, in 50% of instances the null hypothesis is true) and that commonly investigated effects are of small-to-medium size (r = .20), estimated false-positive rates would be close to the nominal (type 1-error) level, which is refreshing. Only if 80% of research hypotheses were incorrect, the share of published false positives would raise to about 20%. However, considering the prevalence of *p*-hacking and other questionable research practices (for which there is various evidence, e.g., Head, Holman, Lanfear, Kahn, & Jennions, 2015; John et al., 2012), actual false-positive rates must be higher.

Studies featuring online samples (vs. conventionally recruited samples) in either journal, and articles providing supplemental files (vs. without) in one journal (PAID), had higher NFs. Similarly, JID, since 2005 solely publishing English-language articles, had a higher NF than prior to its internationalization. It should however be pointed out that these findings are observational and thus confounded with time (publication year). It seems fair to say that various modernization indicators and events (such as the three highlighted here) show relationships with recent trends in the NF; still, causality cannot be established from this correlational approach.

Perhaps surprisingly, and counterintuitively, easily collected convenience samples (comprised of university students) were smaller, instead of larger, than other (general population, community) samples. The prevalence of such surrogate samples, representing a very narrow sampling frame, was high (35–45%) and, what is more, has increased over time. We deem it necessary to discuss and to investigate further this large share of student samples, and time trends therein, coupled with the fact that such easily collected samples on average are smaller than less easily collected non-student samples. In this context, one topic of further inquiry would be to disentangle and clarify whether the limited size of student samples more likely reflects limited student participant pools at smaller departments and institutions ("inevitably"), or else, that, regardless of the actual size of such participant pools, student samples from the outset might be designed with narrower sampling frames and plans and less data-collection effort ("deliberately").

There has been much research interest recently in peculiarities observed in the distributions of p values, as published across empirical research articles (Gerber & Malhotra, 2008; Head et al., 2015; Masicampo & Lalande, 2012; Simonsohn, Nelson, & Simmons, 2014). Our evidence suggests that interesting patterns may also be found in the distribution of published sample size numbers, to which a further unexpected insight from our NF analysis pertains.

Specifically, the distribution of sample sizes overall showed clear multimodality, with all modal values occurring at positions divisible by 10. Added to this, there were surplusses of sample sizes just beyond $N = 100$ and of even-numbered sample sizes in general. All of this is unlikely to occur by chance and anyway hard to explain. It may be argued that planned data collection in experimental designs, along with random assignment of study participants to experimental groups, may lead to even-numbered sample sizes which also are divisible by 10. However, the share of experimental designs in our NF analysis was simply too small to account for these phenomena.

To elucidate such digit preferences for sample sizes, we find it a more plausible idea that researchers, instead of conducting proper sample size planning through effect-magnitude estimates and *a priori* power analyses (the prevalence of which might be less than 10% in currently published psychological research; Kühberger, Fritz, & Scherndl, 2014), apparently rely on certain rules of thumb, traditions, and tastes prevailing in their field and specialty, and perhaps might even be prone to some kind of belief in a "law of nice, round numbers." These preferences for sample size numbers observed here appear worthy of further exploration.

We end on a more general point, in suggesting to promote the NF more widely, with the intention of journal-wide implementation of sample size information in published articles. In our era, where submission, peer-review, and acceptance of scholarly manuscripts take place online, it would be simple and easy to transmit, already upon submission of a paper, NF-relevant information (analysis N, sample type, and type of study design) over a journal's web submission portal, which metadata then could be made visible, accessible, and usable.

Abstracts, summarizing a paper's content, are ubiquitous in scholarly publishing and mandatory for authors to write them up. Keyword lists for scientific articles are similarly widespread and obligatory. In addition, journals now increasingly feature so-called Highlights sections. These are intended as at-a-glance overviews, which distill down the main study features and findings to three to five statements of half-sentence length. Similarly, text boxes like "What is already known on this topic" and "What this study adds," accompanying published journal papers, are on the rise, as are translational abstracts, which are tailor-made to be comprehensible for non-professionals.

In the spirit of the ultra-brief "21-word solution" (Simmons et al., 2012), a proposal that has been presented to disclose study procedural details and to counteract questionable research practices, we therefore put forward a call for communicating and disseminating NF-relevant information in journal articles. Doing so would be as compact and effortless as assembling the keyword list or the Highlights section for one's own article. And surely a good thing.

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., … Perugini, M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*, 108–119. https://doi.org/10.1002/per.1919

Brembs, B., Button, K., & Munafò, M. (2013). Deep impact: Unintended consequences of journal rank. *Frontiers in Human Neuroscience, 7*, 291. https://doi.org/10.3389/fnhum.2013.00291

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145–153. https://doi.org/10.1037/h0045186

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

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

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers, 28*, 1–11. https://doi.org/10.3758/BF03203630

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*, 891–904. https://doi.org/10.1007/s11192-011-0494-7

Fraley, R. C., & Vazire, S. (2014). The *N*-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS One, 9*, e109019. https://doi.org/10.1371/journal.pone.0109019

Gerber, A. S., & Malhotra, N. (2008). Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research, 37*, 3–30. https://doi.org/10.1177/0049124108318973

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology, 79*, 314–316. https://doi.org/10.1037/0021-9010.79.2.314

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of *p*-hacking in science. *PLoS Biology, 13*, e1002106. https://doi.org/10.1371/journal.pbio.1002106

Hennig, J. (2005). Editorial. *Journal of Individual Differences, 26*, 1. https://doi.org/10.1027/1614-0001.26.1.1

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524–532. https://doi.org/10.1177/0956797611430953

Kraemer, H. C., & Andrews, G. (1982). A nonparametric technique for meta-analysis effect size calculation. *Psychological Bulletin, 91*, 404–412. https://doi.org/10.1037/0033-2909.91.2.404

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS One, 9*, e105825. https://doi.org/10.1371/journal.pone.0105825

Lamberink, H. J., Otte, W. M., Sinke, M. R., Lakens, D., Glasziou, P. P., Tijdink, J. K., & Vinkers, C. H. (2018). Statistical power of clinical trials increased while effect size remained stable: An empirical analysis of 136,212 clinical trials between 1975 and 2014. *Journal of Clinical Epidemiology, 102*, 123–128. https://doi.org/10.1016/j.jclinepi.2018.06.014

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of *p* values just below .05. *Quarterly Journal of Experimental Psychology, 65*, 2271–2279. https://doi.org/10.1080/17470218.2012.711335

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716. https://doi.org/10.1126/science.aac4716

Reardon, K. W., Smack, A. J., Herzhoff, K., & Tackett, J. L. (2019). An *N*-pact factor for clinical psychological research. *Journal of Abnormal Psychology, 128*, 493–499. https://doi.org/10.1037/abn0000435

Schweizer, G., & Furley, P. (2016). Reproducible research in sport and exercise psychology: The role of sample sizes. *Psychology of Sport and Exercise, 23*, 114–122. https://doi.org/10.1016/j.psychsport.2015.11.005

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366. https://doi.org/10.1177/0956797611417632

Simmons, J., Nelson, L., & Simonsohn, U. (2012). A 21 word solution. *Dialogue: Official Newsletter of the Society for Personality and Social Psychology, 26*(2), 4–7. https://doi.org/10.2139/ssrn.2160588

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*, 534–547. https://doi.org/10.1037/a0033242

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology, 15*, e2000797. https://doi.org/10.1371/journal.pbio.2000797

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … Bouwman, J. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*, 160018. https://doi.org/10.1038/sdata.2016.18

History

Received November 19, 2018
Revision received June 12, 2019
Accepted June 16, 2019
Published online December 20, 2019

Open Data

All components required for reproducible data analysis are archived at the Open Science Framework (annotated coding form: https://osf.io/kcuyj/; dataset for analysis: https://osf.io/nrmh3/; code to reproduce the analyses: https://osf.io/yvnp6/). R code to reproduce the figures is available at https://osf.io/rbdkz/ (Figure 1), https://osf.io/ayce2/ (Figure 2), https://osf.io/2ktcj/ (Figure 3), and https://osf.io/u6xhb/ (Figure 4).

Martin Voracek

Department of Basic Psychological Research and Research Methods
School of Psychology
University of Vienna
Liebiggasse 5
1010 Vienna
Austria
martin.voracek@univie.ac.at